Multivariate Verfahren 2

discriminant analysis

Helmut Waldl

June 4th and 5th 2012

discriminant analysis

class regions

With each decision rule Ω_x is partitioned in disjoint class regions $D_1, \ldots, D_g$ $\Omega_x = \bigcup_{i=1}^{g} D_i$. The class regions are defined as follows:

$$\mathring{D}_k := \{\mathbf{x} \in \Omega_x | d_k(\mathbf{x}) > d_i(\mathbf{x}), \forall i \neq k\} = D_k \backslash \partial D_k$$

$\mathring{D}_k$ are the inner points of D_k, i.e. D_k without its boundary ∂D_k.

In D_k the discriminant function of the kth class is maximum.

∂D_k are the separation planes between the class regions, on ∂D_k we have: $\exists i \neq k : \; d_k(\mathbf{x}) = d_i(\mathbf{x})$.

Also ∂D_k must be uniquely assigned to a class region so that the partitioning of Ω_x is well-defined.

discriminant analysis

class regions

With each decision rule Ω_x is partitioned in disjoint class regions $D_1, \ldots, D_g$ $\quad \Omega_x = \bigcup_{i=1}^{g} D_i$. The class regions are defined as follows:

$$\mathring{D}_k := \{\mathbf{x} \in \Omega_x | d_k(\mathbf{x}) > d_i(\mathbf{x}), \forall i \neq k\} = D_k \backslash \partial D_k$$

$\mathring{D}_k$ are the inner points of D_k, i.e. D_k without its boundary ∂D_k.

In D_k the discriminant function of the kth class is maximum.

∂D_k are the separation planes between the class regions, on ∂D_k we have: $\exists i \neq k : \; d_k(\mathbf{x}) = d_i(\mathbf{x})$.

Also ∂D_k must be uniquely assigned to a class region so that the partitioning of Ω_x is well-defined.

discriminant analysis

class regions

With each decision rule Ω_x is partitioned in disjoint class regions $D_1, \ldots, D_g$ $\Omega_x = \bigcup_{i=1}^{g} D_i$. The class regions are defined as follows:

$$\mathring{D}_k := \{\mathbf{x} \in \Omega_x | d_k(\mathbf{x}) > d_i(\mathbf{x}), \forall i \neq k\} = D_k \backslash \partial D_k$$

$\mathring{D}_k$ are the inner points of D_k, i.e. D_k without its boundary ∂D_k.

In D_k the discriminant function of the kth class is maximum.

∂D_k are the separation planes between the class regions, on ∂D_k we have: $\exists i \neq k : \; d_k(\mathbf{x}) = d_i(\mathbf{x})$.

Also ∂D_k must be uniquely assigned to a class region so that the partitioning of Ω_x is well-defined.

discriminant analysis

class regions

With each decision rule Ω_x is partitioned in disjoint class regions $D_1, \ldots, D_g$ $\Omega_x = \bigcup_{i=1}^{g} D_i$. The class regions are defined as follows:

$$\mathring{D}_k := \{\mathbf{x} \in \Omega_x | d_k(\mathbf{x}) > d_i(\mathbf{x}), \forall i \neq k\} = D_k \backslash \partial D_k$$

$\mathring{D}_k$ are the inner points of D_k, i.e. D_k without its boundary ∂D_k.

In D_k the discriminant function of the kth class is maximum.

∂D_k are the separation planes between the class regions, on ∂D_k we have: $\exists i \neq k : \; d_k(\mathbf{x}) = d_i(\mathbf{x})$.

Also ∂D_k must be uniquely assigned to a class region so that the partitioning of Ω_x is well-defined.

discriminant analysis

class regions

With each decision rule Ω_x is partitioned in disjoint class regions $D_1, \ldots, D_g$ $\Omega_x = \bigcup_{i=1}^{g} D_i$. The class regions are defined as follows:

$$\mathring{D}_k := \{\mathbf{x} \in \Omega_x | d_k(\mathbf{x}) > d_i(\mathbf{x}), \forall i \neq k\} = D_k \backslash \partial D_k$$

$\mathring{D}_k$ are the inner points of D_k, i.e. D_k without its boundary ∂D_k.

In D_k the discriminant function of the kth class is maximum.

∂D_k are the separation planes between the class regions, on ∂D_k we have: $\exists i \neq k : \; d_k(\mathbf{x}) = d_i(\mathbf{x})$.

Also ∂D_k must be uniquely assigned to a class region so that the partitioning of Ω_x is well-defined.

discriminant analysis

class regions

Why class regions? Instead of the decision function e in many applications the partition of Ω_x induced by the used decision rule is specified. We have:

$$e(\mathbf{x}) = \hat{k} \qquad \Longleftrightarrow \qquad \mathbf{x} \in D_{\hat{k}}$$

If we use class regions we get an easier representation of the error rates.

Example: total error rate, $g = 2$

$$\varepsilon(e) = \varepsilon(D, f) = P(\omega \text{ is misclassified}) = P(e(\mathbf{x}) \neq k) =$$

$$= P(\mathbf{x} \in D_2, k = 1) + P(\mathbf{x} \in D_1, k = 2) =$$

$$= P(\mathbf{x} \in D_2 | k = 1) \cdot p(1) + P(\mathbf{x} \in D_1 | k = 2) \cdot p(2) =$$

$$= \int_{D_2} f(\mathbf{x}|1) \cdot p(1)\, dx + \int_{D_1} f(\mathbf{x}|2) \cdot p(2)\, dx$$

discriminant analysis

class regions

Why class regions? Instead of the decision function e in many applications the partition of Ω_x induced by the used decision rule is specified. We have:

$$e(\mathbf{x}) = \hat{k} \qquad \Longleftrightarrow \qquad \mathbf{x} \in D_{\hat{k}}$$

If we use class regions we get an easier representation of the error rates.

Example: total error rate, $g = 2$

$$\varepsilon(e) = \varepsilon(D, f) = P(\omega \text{ is misclassified}) = P(e(\mathbf{x}) \neq k) =$$

$$= P(\mathbf{x} \in D_2, k = 1) + P(\mathbf{x} \in D_1, k = 2) =$$

$$= P(\mathbf{x} \in D_2 | k = 1) \cdot p(1) + P(\mathbf{x} \in D_1 | k = 2) \cdot p(2) =$$

$$= \int_{D_2} f(\mathbf{x}|1) \cdot p(1)\, dx \; + \; \int_{D_1} f(\mathbf{x}|2) \cdot p(2)\, dx$$

discriminant analysis

class regions

Example: individual error rates;

$$\varepsilon_{k,\hat{k}} = P(\mathbf{x} \in D_{\hat{k}}|k) = \int_{D_{\hat{k}}} f(\mathbf{x}|k)\, dx$$

estimated decision rules and error rates

Up to now $p(k)$, $f(\mathbf{x}|k)$ and hence $p(k|\mathbf{x})$ were assumed to be known. In practice we have to estimate the distributions and their parameters respectively.

In most cases we assume certain probability distributions and then estimate the according parameters, i.e. we assess the distribution in parametric form: $f(\mathbf{x}|\theta_k)$, $p(k|\mathbf{x}, \theta)$ or even the discriminant function $d_k(\mathbf{x}|\theta_k)$

discriminant analysis

class regions

Example: individual error rates;

$$\varepsilon_{k,\hat{k}} = P(\mathbf{x} \in D_{\hat{k}}|k) = \int_{D_{\hat{k}}} f(\mathbf{x}|k)\, dx$$

estimated decision rules and error rates

Up to now $p(k)$, $f(\mathbf{x}|k)$ and hence $p(k|\mathbf{x})$ were assumed to be known. In practice we have to estimate the distributions and their parameters respectively.

In most cases we assume certain probability distributions and then estimate the according parameters, i.e. we assess the distribution in parametric form: $f(\mathbf{x}|\theta_k)$, $p(k|\mathbf{x}, \theta)$ or even the discriminant function $d_k(\mathbf{x}|\theta_k)$

discriminant analysis

class regions

Example: individual error rates;

$$\varepsilon_{k,\hat{k}} = P(\mathbf{x} \in D_{\hat{k}}|k) = \int_{D_{\hat{k}}} f(\mathbf{x}|k)\, dx$$

estimated decision rules and error rates

Up to now $p(k)$, $f(\mathbf{x}|k)$ and hence $p(k|\mathbf{x})$ were assumed to be known. In practice we have to estimate the distributions and their parameters respectively.

In most cases we assume certain probability distributions and then estimate the according parameters, i.e. we assess the distribution in parametric form: $f(\mathbf{x}|\theta_k)$, $p(k|\mathbf{x}, \theta)$ or even the discriminant function $d_k(\mathbf{x}|\theta_k)$

discriminant analysis

class regions

Example: individual error rates;

$$\varepsilon_{k,\hat{k}} = P(\mathbf{x} \in D_{\hat{k}}|k) = \int_{D_{\hat{k}}} f(\mathbf{x}|k)\, dx$$

estimated decision rules and error rates

Up to now $p(k)$, $f(\mathbf{x}|k)$ and hence $p(k|\mathbf{x})$ were assumed to be known. In practice we have to estimate the distributions and their parameters respectively.

In most cases we assume certain probability distributions and then estimate the according parameters, i.e. we assess the distribution in parametric form: $f(\mathbf{x}|\theta_k)$, $p(k|\mathbf{x}, \theta)$ or even the discriminant function $d_k(\mathbf{x}|\theta_k)$

discriminant analysis

estimated decision rules and error rates

estimation method: e.g. ML- or LS-estimation - there are no limitations on the used method. Estimation only with a learning sample.

The estimated decision rules depend on the estimation- and sampling-method:

ML estimation, unrestricted random sample (total sample)

$(\mathbf{x}_i, k_i)$ $i = 1, \ldots N$ are independent observations from the distribution of $(\mathbf{x}, k)$.

The likelihood function with given a-priori probabilities $p(k)$ is

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i | \theta_{k_i})$$

discriminant analysis

estimated decision rules and error rates

estimation method: e.g. ML- or LS-estimation - there are no limitations on the used method. Estimation only with a learning sample.

The estimated decision rules depend on the estimation- and sampling-method:

ML estimation, unrestricted random sample (total sample)

$(\mathbf{x}_i, k_i)$ $i = 1, \ldots N$ are independent observations from the distribution of $(\mathbf{x}, k)$.

The likelihood function with given a-priori probabilities $p(k)$ is

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i | \theta_{k_i})$$

discriminant analysis

estimated decision rules and error rates

estimation method: e.g. ML- or LS-estimation - there are no limitations on the used method. Estimation only with a learning sample.

The estimated decision rules depend on the estimation- and sampling-method:

ML estimation, unrestricted random sample (total sample)

$(\mathbf{x}_i, k_i)$ $i = 1, \ldots N$ are independent observations from the distribution of $(\mathbf{x}, k)$.

The likelihood function with given a-priori probabilities $p(k)$ is

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i | \theta_{k_i})$$

discriminant analysis

estimated decision rules and error rates

estimation method: e.g. ML- or LS-estimation - there are no limitations on the used method. Estimation only with a learning sample.

The estimated decision rules depend on the estimation- and sampling-method:

ML estimation, unrestricted random sample (total sample)

$(\mathbf{x}_i, k_i)$ $i = 1, \ldots N$ are independent observations from the distribution of $(\mathbf{x}, k)$.

The likelihood function with given a-priori probabilities $p(k)$ is

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i | \theta_{k_i})$$

discriminant analysis

estimated decision rules and error rates

estimation method: e.g. ML- or LS-estimation - there are no limitations on the used method. Estimation only with a learning sample.

The estimated decision rules depend on the estimation- and sampling-method:

ML estimation, unrestricted random sample (total sample)

$(\mathbf{x}_i, k_i)$ $i = 1, \ldots N$ are independent observations from the distribution of $(\mathbf{x}, k)$.

The likelihood function with given a-priori probabilities $p(k)$ is

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i | \theta_{k_i})$$

discriminant analysis

ML estimation, unrestricted random sample (total sample)

We consider L_T as a function of $(\theta_1, \ldots, \theta_g, p(k))$ and maximize it. That yields the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_g, \hat{p}(k) = \frac{N_k}{N}$, where N_k is the number of observations in class k.

The discriminant function of the Bayes decision rule $d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k)$ is then replaced by the estimated discriminant function:

$$d_k(\mathbf{x}|\hat{\theta}_k) = \hat{p}(k) \cdot f(\mathbf{x}|\hat{\theta}_k)$$

If $p(k)$ is known we will of course not use $\hat{p}(k)$.

Analogously we get the estimated ML-discriminant function:

$$d_k(\mathbf{x}|\hat{\theta}_k) = f(\mathbf{x}|\hat{\theta}_k)$$

discriminant analysis

ML estimation, unrestricted random sample (total sample)

We consider L_T as a function of $(\theta_1, \ldots, \theta_g, p(k))$ and maximize it. That yields the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_g, \hat{p}(k) = \frac{N_k}{N}$, where N_k is the number of observations in class k.

The discriminant function of the Bayes decision rule $d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k)$ is then replaced by the estimated discriminant function:

$$d_k(\mathbf{x}|\hat{\theta}_k) = \hat{p}(k) \cdot f(\mathbf{x}|\hat{\theta}_k)$$

If $p(k)$ is known we will of course not use $\hat{p}(k)$.

Analogously we get the estimated ML-discriminant function:

$$d_k(\mathbf{x}|\hat{\theta}_k) = f(\mathbf{x}|\hat{\theta}_k)$$

discriminant analysis

ML estimation, unrestricted random sample (total sample)

We consider L_T as a function of $(\theta_1, \ldots, \theta_g, p(k))$ and maximize it. That yields the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_g, \hat{p}(k) = \frac{N_k}{N}$, where N_k is the number of observations in class k.

The discriminant function of the Bayes decision rule $d_k(\mathbf{x}) = p(k) \cdot f(\mathbf{x}|k)$ is then replaced by the estimated discriminant function:

$$d_k(\mathbf{x}|\hat{\theta}_k) = \hat{p}(k) \cdot f(\mathbf{x}|\hat{\theta}_k)$$

If $p(k)$ is known we will of course not use $\hat{p}(k)$.

Analogously we get the estimated ML-discriminant function:

$$d_k(\mathbf{x}|\hat{\theta}_k) = f(\mathbf{x}|\hat{\theta}_k)$$

discriminant analysis

ML estimation, unrestricted random sample (total sample)

The likelihood function can also be computed for a given a posteriori probability:

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i|\mathbf{x}_i, \theta_{k_i}) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i)$$

The estimated Bayes discriminant function then is:

$$d_k(\mathbf{x}|\hat{\theta}_k) = \hat{p}(k|\mathbf{x}, \hat{\theta}_k)$$

Caution: Also the mixture distribution $f(\mathbf{x})$ may contain information about the parameter θ_k. Do not only maximize the first factor!

discriminant analysis

ML estimation, unrestricted random sample (total sample)

The likelihood function can also be computed for a given a posteriori probability:

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i|\mathbf{x}_i, \theta_{k_i}) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i)$$

The estimated Bayes discriminant function then is:

$$d_k(\mathbf{x}|\hat{\theta}_k) = \hat{p}(k|\mathbf{x}, \hat{\theta}_k)$$

Caution: Also the mixture distribution $f(\mathbf{x})$ may contain information about the parameter θ_k. Do not only maximize the first factor!

discriminant analysis

ML estimation, unrestricted random sample (total sample)

The likelihood function can also be computed for a given a posteriori probability:

$$L_T = \prod_{i=1}^{N} f(\mathbf{x}_i, k_i) = \prod_{i=1}^{N} p(k_i|\mathbf{x}_i, \theta_{k_i}) \cdot \prod_{i=1}^{N} f(\mathbf{x}_i)$$

The estimated Bayes discriminant function then is:

$$d_k(\mathbf{x}|\hat{\theta}_k) = \hat{p}(k|\mathbf{x}, \hat{\theta}_k)$$

Caution: Also the mixture distribution $f(\mathbf{x})$ may contain information about the parameter θ_k. Do not only maximize the first factor!

discriminant analysis

ML estimation, stratified by class sampling

From each of the g classes N_k (fixed) observations of $\mathbf{x}$ are drawn ($\rightarrow f(\mathbf{x}|k)$). This is necessary if some classes are very small, estimates for these classes would be inaccurate also with big samples.

With the likelihood function $L_k = \prod_{i=1}^{N} f(\mathbf{x}_i|\theta_{k_i})$ we get the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_g$. The a priori distribution $p(k)$ cannot be estimated because we fixed N_k. Hence there exists only an estimated ML discriminant function.

We also have difficulties in the ML estimation of the a posteriori probabilities $p(k|\mathbf{x}, \theta_\mathbf{k})$.

discriminant analysis

ML estimation, stratified by class sampling

From each of the g classes N_k (fixed) observations of $\mathbf{x}$ are drawn ($\rightarrow f(\mathbf{x}|k)$). This is necessary if some classes are very small, estimates for these classes would be inaccurate also with big samples.

With the likelihood function $L_k = \prod_{i=1}^{N} f(\mathbf{x}_i|\theta_{k_i})$ we get the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_g$. The a priori distribution $p(k)$ cannot be estimated because we fixed N_k. Hence there exists only an estimated ML discriminant function.

We also have difficulties in the ML estimation of the a posteriori probabilities $p(k|\mathbf{x}, \theta_\mathbf{k})$.

discriminant analysis

ML estimation, stratified by class sampling

From each of the g classes N_k (fixed) observations of $\mathbf{x}$ are drawn ($\rightarrow f(\mathbf{x}|k)$). This is necessary if some classes are very small, estimates for these classes would be inaccurate also with big samples.

With the likelihood function $L_k = \prod_{i=1}^{N} f(\mathbf{x}_i|\theta_{k_i})$ we get the estimates $\hat{\theta}_1, \ldots, \hat{\theta}_g$. The a priori distribution $p(k)$ cannot be estimated because we fixed N_k. Hence there exists only an estimated ML discriminant function.

We also have difficulties in the ML estimation of the a posteriori probabilities $p(k|\mathbf{x}, \theta_\mathbf{k})$.

discriminant analysis

ML estimation, stratified by **x**-values sampling

For N given values $\mathbf{x}_1, \ldots, \mathbf{x}_N$ the class index is independently observed as $k|\mathbf{x}_1, \ldots, k|\mathbf{x}_N$ (e.g. systematic experiments in medicine).

The likelihood function is self-evidently parametrized using the a posteriori distribution

$$L_x = \prod_{i=1}^{N} p(k|\mathbf{x}_i) = \prod_{i=1}^{N} p(k_i|\mathbf{x}_i, \theta_{k_i})$$

If the mixture distribution $f(\mathbf{x})$ contains no information about the parameter θ we get the same estimates as with unrestricted random sampling because

$$L_T = L_x \cdot \prod_{i=1}^{N} f(\mathbf{x}_i)$$

discriminant analysis

ML estimation, stratified by **x**-values sampling

For N given values $\mathbf{x}_1, \ldots, \mathbf{x}_N$ the class index is independently observed as $k|\mathbf{x}_1, \ldots, k|\mathbf{x}_N$ (e.g. systematic experiments in medicine).

The likelihood function is self-evidently parametrized using the a posteriori distribution

$$L_x = \prod_{i=1}^{N} p(k|\mathbf{x}_i) = \prod_{i=1}^{N} p(k_i|\mathbf{x}_i, \theta_{k_i})$$

If the mixture distribution $f(\mathbf{x})$ contains no information about the parameter θ we get the same estimates as with unrestricted random sampling because

$$L_T = L_x \cdot \prod_{i=1}^{N} f(\mathbf{x}_i)$$

discriminant analysis

ML estimation, stratified by **x**-values sampling

For N given values $\mathbf{x}_1, \ldots, \mathbf{x}_N$ the class index is independently observed as $k|\mathbf{x}_1, \ldots, k|\mathbf{x}_N$ (e.g. systematic experiments in medicine).

The likelihood function is self-evidently parametrized using the a posteriori distribution

$$L_x = \prod_{i=1}^{N} p(k|\mathbf{x}_i) = \prod_{i=1}^{N} p(k_i|\mathbf{x}_i, \theta_{k_i})$$

If the mixture distribution $f(\mathbf{x})$ contains no information about the parameter θ we get the same estimates as with unrestricted random sampling because

$$L_T = L_x \cdot \prod_{i=1}^{N} f(\mathbf{x}_i)$$

discriminant analysis

estimated error rates

Each theoretical decision rule implies a partitioning in class regions $D = (D_1, \ldots, D_g)$. Changing to estimated decision rules we have to use estimated error rates instead of the theoretical error rate $\varepsilon(D, f)$.

We will demonstrate the changeover for $g = 2$, a generalization for arbitrary g is an easy exercise.

theoretical error rate:

$$\varepsilon(D, f) = p(1) \cdot \underbrace{\int_{D_2} f(\mathbf{x}|1)\, dx}_{\varepsilon_{1\,2}(D,f)} + p(2) \cdot \underbrace{\int_{D_1} f(\mathbf{x}|2)\, dx}_{\varepsilon_{2\,1}(D,f)}$$

individual error rates

This error rate is minimum if we use the Bayes decision rule. But now we have estimated decision rules $\hat{D} = (\hat{D}_1, \hat{D}_2)$, i.e. we have to compute the actual error rate.

discriminant analysis

estimated error rates

Each theoretical decision rule implies a partitioning in class regions $D = (D_1, \ldots, D_g)$. Changing to estimated decision rules we have to use estimated error rates instead of the theoretical error rate $\varepsilon(D, f)$.

We will demonstrate the changeover for $g = 2$, a generalization for arbitrary g is an easy exercise.

theoretical error rate:

$$\varepsilon(D, f) = p(1) \cdot \underbrace{\int_{D_2} f(\mathbf{x}|1)\, dx}_{\varepsilon_{1\,2}(D,f)} + p(2) \cdot \underbrace{\int_{D_1} f(\mathbf{x}|2)\, dx}_{\varepsilon_{2\,1}(D,f)}$$

individual error rates

This error rate is minimum if we use the Bayes decision rule. But now we have estimated decision rules $\hat{D} = (\hat{D}_1, \hat{D}_2)$, i.e. we have to compute the actual error rate.

discriminant analysis

estimated error rates

Each theoretical decision rule implies a partitioning in class regions $D = (D_1, \ldots, D_g)$. Changing to estimated decision rules we have to use estimated error rates instead of the theoretical error rate $\varepsilon(D, f)$.

We will demonstrate the changeover for $g = 2$, a generalization for arbitrary g is an easy exercise.

theoretical error rate:

$$\varepsilon(D, f) = p(1) \cdot \underbrace{\int_{D_2} f(\mathbf{x}|1)\, dx}_{\varepsilon_{1\,2}(D,f)} + p(2) \cdot \underbrace{\int_{D_1} f(\mathbf{x}|2)\, dx}_{\varepsilon_{2\,1}(D,f)}$$

individual error rates

This error rate is minimum if we use the Bayes decision rule. But now we have estimated decision rules $\hat{D} = (\hat{D}_1, \hat{D}_2)$, i.e. we have to compute the actual error rate.

discriminant analysis

estimated error rates

Each theoretical decision rule implies a partitioning in class regions $D = (D_1, \ldots, D_g)$. Changing to estimated decision rules we have to use estimated error rates instead of the theoretical error rate $\varepsilon(D, f)$.

We will demonstrate the changeover for $g = 2$, a generalization for arbitrary g is an easy exercise.

theoretical error rate:

$$\varepsilon(D, f) = p(1) \cdot \underbrace{\int_{D_2} f(\mathbf{x}|1)\, dx}_{\varepsilon_{1\,2}(D,f)} + p(2) \cdot \underbrace{\int_{D_1} f(\mathbf{x}|2)\, dx}_{\varepsilon_{2\,1}(D,f)}$$

individual error rates

This error rate is minimum if we use the Bayes decision rule. But now we have estimated decision rules $\hat{D} = (\hat{D}_1, \hat{D}_2)$, i.e. we have to compute the actual error rate.

discriminant analysis

estimated error rates

actual error rate:

$$\varepsilon(\hat{D}, f) = p(1) \cdot \int_{\hat{D}_2} f(\mathbf{x}|1)\, dx + p(2) \cdot \int_{\hat{D}_1} f(\mathbf{x}|2)\, dx$$

$\varepsilon(\hat{D}, f)$ is a random variate, i.e. in practice we are interested in the expectation:

expected actual error rate: $E(\varepsilon(\hat{D}, f))$

The expectation is computed with the random variates $(k_i, \mathbf{x}_i)$ of the learning sample.

That is all still theoretical, in practice we need an estimate for the actual error rate. A plug-in estimate makes sense also here (We substitute the estimated distribution for the unknown real distribution).

discriminant analysis

estimated error rates

actual error rate:

$$\varepsilon(\hat{D}, f) = p(1) \cdot \int_{\hat{D}_2} f(\mathbf{x}|1)\, dx + p(2) \cdot \int_{\hat{D}_1} f(\mathbf{x}|2)\, dx$$

$\varepsilon(\hat{D}, f)$ is a random variate, i.e. in practice we are interested in the expectation:

expected actual error rate: $E(\varepsilon(\hat{D}, f))$

The expectation is computed with the random variates $(k_i, \mathbf{x}_i)$ of the learning sample.

That is all still theoretical, in practice we need an estimate for the actual error rate. A plug-in estimate makes sense also here (We substitute the estimated distribution for the unknown real distribution).

discriminant analysis

estimated error rates

actual error rate:

$$\varepsilon(\hat{D}, f) = p(1) \cdot \int_{\hat{D}_2} f(\mathbf{x}|1)\, dx + p(2) \cdot \int_{\hat{D}_1} f(\mathbf{x}|2)\, dx$$

$\varepsilon(\hat{D}, f)$ is a random variate, i.e. in practice we are interested in the expectation:

expected actual error rate: $E(\varepsilon(\hat{D}, f))$

The expectation is computed with the random variates $(k_i, \mathbf{x}_i)$ of the learning sample.

That is all still theoretical, in practice we need an estimate for the actual error rate. A plug-in estimate makes sense also here (We substitute the estimated distribution for the unknown real distribution).

discriminant analysis

estimated error rates

actual error rate:

$$\varepsilon(\hat{D}, f) = p(1) \cdot \int_{\hat{D}_2} f(\mathbf{x}|1)\, dx + p(2) \cdot \int_{\hat{D}_1} f(\mathbf{x}|2)\, dx$$

$\varepsilon(\hat{D}, f)$ is a random variate, i.e. in practice we are interested in the expectation:

expected actual error rate: $E(\varepsilon(\hat{D}, f))$

The expectation is computed with the random variates $(k_i, \mathbf{x}_i)$ of the learning sample.

That is all still theoretical, in practice we need an estimate for the actual error rate. A plug-in estimate makes sense also here (We substitute the estimated distribution for the unknown real distribution).

discriminant analysis

estimated error rates

estimated actual error rate:

$$\varepsilon(\hat{D}, \hat{f}) = \hat{p}(1) \cdot \int_{\hat{D}_2} \hat{f}(\mathbf{x}|1)\ dx + \hat{p}(2) \cdot \int_{\hat{D}_1} \hat{f}(\mathbf{x}|2)\ dx$$

If $\hat{f}(\mathbf{x}|k)$ is an unbiased estimate we get for the Bayes decision rule:

$$E(\varepsilon(\hat{D}, \hat{f})) \leq \varepsilon(D, f) \leq E(\varepsilon(\hat{D}, f))$$

discriminant analysis

estimated error rates

estimated actual error rate:

$$\varepsilon(\hat{D}, \hat{f}) = \hat{p}(1) \cdot \int_{\hat{D}_2} \hat{f}(\mathbf{x}|1)\ dx + \hat{p}(2) \cdot \int_{\hat{D}_1} \hat{f}(\mathbf{x}|2)\ dx$$

If $\hat{f}(\mathbf{x}|k)$ is an unbiased estimate we get for the Bayes decision rule:

$$E(\varepsilon(\hat{D}, \hat{f})) \leq \varepsilon(D, f) \leq E(\varepsilon(\hat{D}, f))$$

discriminant analysis

convergence of estimated discriminant function and actual error rate

Theorem: If we have $\hat{f}(\mathbf{x}|k) \overset{N\to\infty}{\longrightarrow} f(\mathbf{x}|k)$ and that for all k with positive a priori probability $p(k)$,
then also the estimated discriminant function converges to the optimal Bayes discriminant function:

$$\hat{p}(k) \cdot \hat{f}(\mathbf{x}|k) \overset{N\to\infty}{\longrightarrow} p(k) \cdot f(\mathbf{x}|k) \qquad k = 1, \ldots, g$$

where $\hat{p}(k) = \frac{N_k}{N}$. If furthermore

$$\int_{\Omega_x} \sum_{k=1}^{g} \hat{p}(k) \cdot \hat{f}(\mathbf{x}|k)\, dx \longrightarrow 1$$

(that is always true for parametric estimation $\hat{f}(\mathbf{x}|k) = f(\mathbf{x}|\hat{\theta}_k)$ because $f(\mathbf{x}|\hat{\theta}_k)$ is a pdf),
then also the actual error rate converges:

$$\varepsilon(\hat{D}, f) \longrightarrow \varepsilon(D, f)$$

discriminant analysis

convergence of estimated discriminant function and actual error rate

Theorem: If we have $\hat{f}(\mathbf{x}|k) \overset{N\to\infty}{\longrightarrow} f(\mathbf{x}|k)$ and that for all k with positive a priori probability $p(k)$,
then also the estimated discriminant function converges to the optimal Bayes discriminant function:

$$\hat{p}(k) \cdot \hat{f}(\mathbf{x}|k) \overset{N\to\infty}{\longrightarrow} p(k) \cdot f(\mathbf{x}|k) \qquad k = 1, \ldots, g$$

where $\hat{p}(k) = \frac{N_k}{N}$. If furthermore

$$\int_{\Omega_x} \sum_{k=1}^{g} \hat{p}(k) \cdot \hat{f}(\mathbf{x}|k)\, dx \longrightarrow 1$$

(that is always true for parametric estimation $\hat{f}(\mathbf{x}|k) = f(\mathbf{x}|\hat{\theta}_k)$ because $f(\mathbf{x}|\hat{\theta}_k)$ is a pdf),
then also the actual error rate converges:

$$\varepsilon(\hat{D}, f) \longrightarrow \varepsilon(D, f)$$

discriminant analysis

convergence of estimated discriminant function and actual error rate

Remark:

With parametric estimation $f(\mathbf{x}|\hat{\theta}_k)$ the estimate $\hat{f}(\mathbf{x}|k)$ converges if the estimate $\hat{\theta}_k$ is consistent and if $f(\mathbf{x}|\theta_k)$ is continuous in θ_k.

discriminant analysis

special case: normal distributed variates - classical discriminant analysis

assumption: $(\mathbf{x}|k) \sim \mathbf{N}(\mu_k; \Sigma_k)$ ($\mathbf{x}$ is p-dimensional)

If we use the Bayes decision rule ($p(k) \cdot f(\mathbf{x}|k) \rightarrow \max!$) we get the logarithmic discriminant function

$$d_k(\mathbf{x}) = \ln(p(k)) + \ln(f(\mathbf{x}|k)) =$$

$$= \ln(p(k)) \underbrace{-\frac{p}{2}\ln(2\pi)}_{\text{does not affect maximizing}} - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)$$

We get the discriminant function of the ML decision rule if we omit the a priori pdf $\ln(p(k))$.

discriminant analysis

special case: normal distributed variates - classical discriminant analysis

assumption: $(\mathbf{x}|k) \sim \mathbf{N}(\mu_k; \Sigma_k)$ ($\mathbf{x}$ is p-dimensional)

If we use the Bayes decision rule ($p(k) \cdot f(\mathbf{x}|k) \rightarrow \max!$) we get the logarithmic discriminant function

$$d_k(\mathbf{x}) = \ln(p(k)) + \ln(f(\mathbf{x}|k)) =$$

$$= \ln(p(k)) \underbrace{-\frac{p}{2}\ln(2\pi)}_{\text{does not affect maximizing}} - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)$$

We get the discriminant function of the ML decision rule if we omit the a priori pdf $\ln(p(k))$.

discriminant analysis

special case: normal distributed variates - classical discriminant analysis

assumption: $(\mathbf{x}|k) \sim \mathbf{N}(\mu_k; \Sigma_k)$ ($\mathbf{x}$ is p-dimensional)

If we use the Bayes decision rule ($p(k) \cdot f(\mathbf{x}|k) \rightarrow \max!$) we get the logarithmic discriminant function

$$d_k(\mathbf{x}) = \ln(p(k)) + \ln(f(\mathbf{x}|k)) =$$

$$= \ln(p(k)) \underbrace{-\frac{p}{2}\ln(2\pi)}_{\text{does not affect maximizing}} - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)$$

We get the discriminant function of the ML decision rule if we omit the a priori pdf $\ln(p(k))$.

discriminant analysis

special case: independent homoscedastic, normal distributed variates: $\Sigma_k = \sigma^2 I$

We have: $|\Sigma_k| = \sigma^{2p}$ $\quad$ $\Sigma_k^{-1} = \frac{1}{\sigma} I$ $\quad$ and thus

$$d_k(\mathbf{x}) = \ln(p(k)) - \frac{1}{2}\ln(\sigma^{2p}) - \frac{1}{2\sigma^2}(\mathbf{x} - \mu_k)^T(\mathbf{x} - \mu_k) =$$

$$= \ln(p(k)) \underbrace{-p\ln(\sigma)}_{\text{does not affect maximizing}} - \frac{\|\mathbf{x} - \mu_k\|^2}{2 \cdot \sigma^2}$$

If the a priori probabilities are equal or if we don't know them we get the ML discriminant function

$$d_k(\mathbf{x}) = -\|\mathbf{x} - \mu_k\|^2$$

i.e. ω is assigned to the class k whose center μ_k has the smallest Euclidian distance to $\mathbf{x}$ $\quad \longrightarrow$ minimum distance classification.

discriminant analysis

special case: independent homoscedastic, normal distributed variates: $\Sigma_k = \sigma^2 I$

We have: $|\Sigma_k| = \sigma^{2p}$ $\quad$ $\Sigma_k^{-1} = \frac{1}{\sigma} I$ $\quad$ and thus

$$d_k(\mathbf{x}) = \ln(p(k)) - \frac{1}{2}\ln(\sigma^{2p}) - \frac{1}{2\sigma^2}(\mathbf{x} - \mu_k)^T(\mathbf{x} - \mu_k) =$$

$$= \ln(p(k)) \underbrace{-p\ln(\sigma)}_{\substack{\text{does not affect} \\ \text{maximizing}}} - \frac{\|\mathbf{x} - \mu_k\|^2}{2 \cdot \sigma^2}$$

If the a priori probabilities are equal or if we don't know them we get the ML discriminant function

$$d_k(\mathbf{x}) = -\|\mathbf{x} - \mu_k\|^2$$

i.e. ω is assigned to the class k whose center μ_k has the smallest Euclidian distance to $\mathbf{x}$ $\quad \longrightarrow$ minimum distance classification.

discriminant analysis

special case: independent homoscedastic, normal distributed variates: $\Sigma_k = \sigma^2 I$

The Bayes discriminant function in fact is linear in $\mathbf{x}$

$$d_k(\mathbf{x}) = \ln(p(k)) - \frac{1}{2\sigma^2}\left(\|\mathbf{x}\|^2 - 2\mu_k^T\mathbf{x} + \|\mu_k\|^2\right) \qquad \text{thus}$$

$$d_k(\mathbf{x}) \;=\; \frac{\mu_k^T}{\sigma^2}\mathbf{x} + \ln(p(k)) - \frac{\|\mu_k\|^2}{2\sigma^2} \;=\; a_k^T\mathbf{x} + a_{k0}$$

The assumption $\Sigma_k = \sigma^2 I$ is very restrictive, the next special case is more general.

discriminant analysis

special case: independent homoscedastic, normal distributed variates: $\Sigma_k = \sigma^2 I$

The Bayes discriminant function in fact is linear in $\mathbf{x}$

$$d_k(\mathbf{x}) = \ln(p(k)) - \frac{1}{2\sigma^2}\left(\|\mathbf{x}\|^2 - 2\mu_k^T\mathbf{x} + \|\mu_k\|^2\right) \qquad \text{thus}$$

$$d_k(\mathbf{x}) \;=\; \frac{\mu_k^T}{\sigma^2}\mathbf{x} + \ln(p(k)) - \frac{\|\mu_k\|^2}{2\sigma^2} \;=\; a_k^T\mathbf{x} + a_{k0}$$

The assumption $\Sigma_k = \sigma^2 I$ is very restrictive, the next special case is more general.

discriminant analysis

special case: normal distributed variates, class-wise identical covariance matrices: $\Sigma_k = \Sigma$

$$d_k(\mathbf{x}) = \ln(p(k)) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}\underbrace{(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)}_{\text{Mahalanobis distance between } \mathbf{x} \text{ and } \mu_k}$$

Again the discriminant function is in fact linear in $\mathbf{x}$:

$$d_k(\mathbf{x}) = \mu_k^T\Sigma^{-1}\mathbf{x} + \ln(p(k)) - \frac{1}{2}\underbrace{\mu_k^T\Sigma^{-1}\mu_k}_{=\|\mu_k\|^2_{\Sigma^{-1}}}$$

discriminant analysis

special case: normal distributed variates, class-wise identical covariance matrices: $\Sigma_k = \Sigma$

$$d_k(\mathbf{x}) = \ln(p(k)) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}\underbrace{(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)}_{\text{Mahalanobis distance between } \mathbf{x} \text{ and } \mu_k}$$

Again the discriminant function is in fact linear in $\mathbf{x}$:

$$d_k(\mathbf{x}) = \mu_k^T\Sigma^{-1}\mathbf{x} + \ln(p(k)) - \frac{1}{2}\underbrace{\mu_k^T\Sigma^{-1}\mu_k}_{=\|\mu_k\|^2_{\Sigma^{-1}}}$$

discriminant analysis

special case: normal distributed variates, general covariance matrices: Σ_k

Only with class-wise different covariance matrices the discriminant function is quadratic in $\mathbf{x}$:

$$d_k(\mathbf{x}) = \mathbf{x}^T A_k \mathbf{x} + a_k^T \mathbf{x} + a_{k0}$$

with $A_k = -\frac{1}{2} \cdot \Sigma_k^{-1}$, $a_k = \Sigma_k^{-1} \mu_k$ and
$a_{k0} = \ln(p(k)) - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \ln |\Sigma_k|$

In all established statistics software packages a linear discriminant analysis is performed by default, i.e. equal covariance matrices are assumed:
$\mathbf{x} \sim \mathbf{N}(\mu_k, \Sigma)$

discriminant analysis

special case: normal distributed variates, general covariance matrices: Σ_k

Only with class-wise different covariance matrices the discriminant function is quadratic in $\mathbf{x}$:

$$d_k(\mathbf{x}) = \mathbf{x}^T A_k \mathbf{x} + a_k^T \mathbf{x} + a_{k0}$$

with $A_k = -\frac{1}{2} \cdot \Sigma_k^{-1}$, $a_k = \Sigma_k^{-1} \mu_k$ and
$a_{k0} = \ln(p(k)) - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \ln |\Sigma_k|$

In all established statistics software packages a linear discriminant analysis is performed by default, i.e. equal covariance matrices are assumed:
$\mathbf{x} \sim \mathbf{N}(\mu_k, \Sigma)$

discriminant analysis

estimated discriminant functions

How do we get the estimated discriminant functions?
Just plug in the unbiased estimates:

$$\bar{\mathbf{x}}_k \text{ for } \mu_k\,, \quad k = 1, \ldots, g \quad \text{and}$$

$$S = \frac{1}{N-g} \sum_{k=1}^{g} \sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T \qquad \text{for } \Sigma$$

$p(k)$ is again estimated by $\frac{N_k}{N}$

$$\Longrightarrow \quad \hat{d}_k(\mathbf{x}) = \bar{\mathbf{x}}_k^T S^{-1} \mathbf{x} - \frac{1}{2} \bar{\mathbf{x}}_k^T S^{-1} \bar{\mathbf{x}}_k + ln(N_k) - \ln N$$

discriminant analysis

estimated discriminant functions

Special case: $g = 2$ classes:

Object ω is assigned to class 1 if

$$\left(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right)^T \cdot a > \ln \frac{p(2)}{p(1)}$$

with $a = S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

If the a priori distribution is unknown or if we want to use the ML decision rule instead of the Bayes decision rule we have to set $\ln \frac{p(k)}{p(i)} = 0$.

discriminant analysis

estimated discriminant functions

Special case: $g = 2$ classes:

Object ω is assigned to class 1 if

$$\left(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right)^T \cdot a > \ln \frac{p(2)}{p(1)}$$

with $a = S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

If the a priori distribution is unknown or if we want to use the ML decision rule instead of the Bayes decision rule we have to set $\ln \frac{p(k)}{p(i)} = 0$.

discriminant analysis

nonparametric ansatz by Fisher

$\mathbf{x} = (x_1, \ldots, x_p)^T$

Idea: transform the p-dimensional problem to a one-dimensional problem.

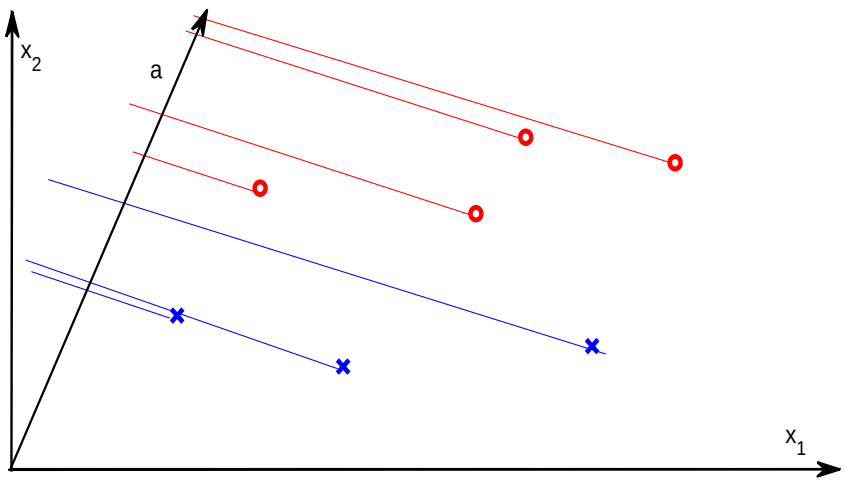
How? With a linear combination of the vector $\mathbf{x}$:

$y = a^T \mathbf{x}\,,\ a^T = (a_1 \ldots a_p)$

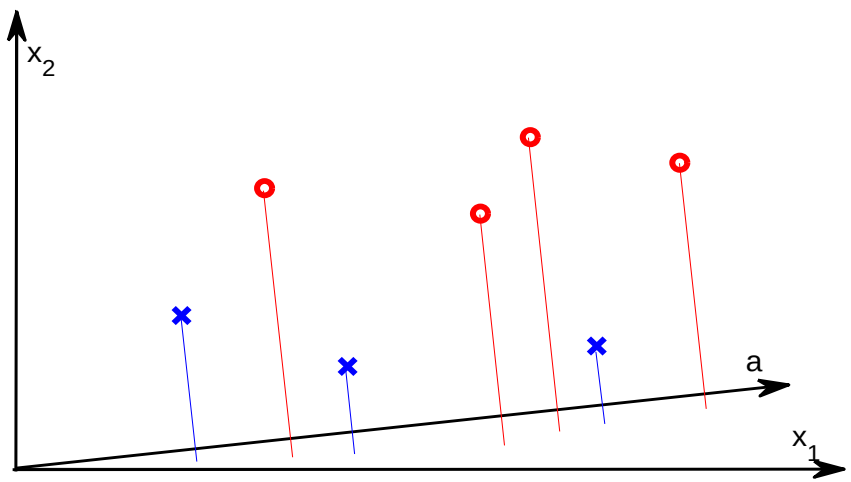
i.e. $\mathbf{x}_{ki}\ , i = 1, \ldots, N_k$ is transformed to $y_{ki} = a^T \mathbf{x}_{ki}$.

With $\|a\| = 1$ the linear combination $a^T \mathbf{x}$ is the projection of the data $\mathbf{x}$ on a straight line with direction a .

Example: $p = 2,\ g = 2$



good choice of a



bad choice of a

discriminant analysis

nonparametric ansatz by Fisher

$\mathbf{x} = (x_1, \ldots, x_p)^T$

Idea: transform the p-dimensional problem to a one-dimensional problem.

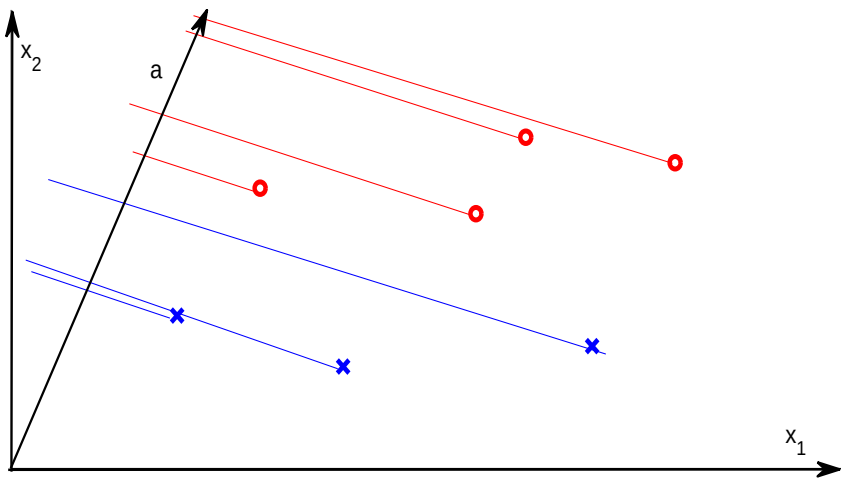
How? With a linear combination of the vector $\mathbf{x}$:

$y = a^T\mathbf{x}\,,\ a^T = (a_1 \ldots a_p)$

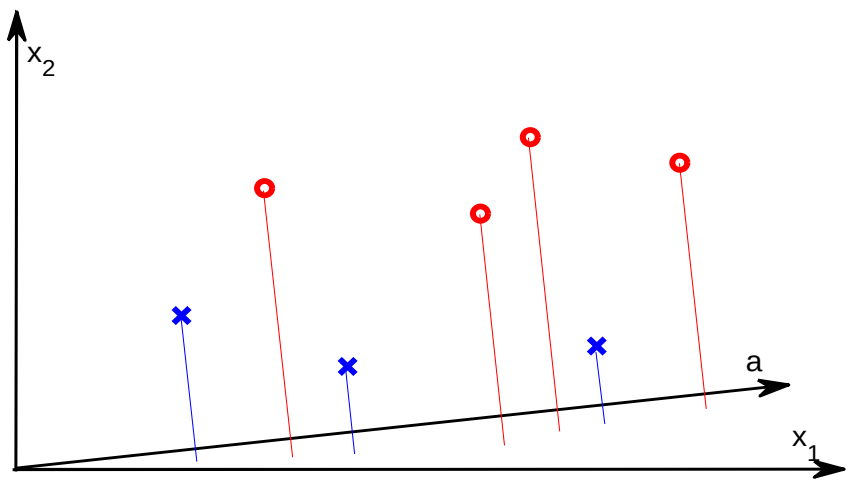
i.e. $\mathbf{x}_{ki}\ , i = 1, \ldots, N_k$ is transformed to $y_{ki} = a^T\mathbf{x}_{ki}$.

With $\|a\| = 1$ the linear combination $a^T\mathbf{x}$ is the projection of the data $\mathbf{x}$ on a straight line with direction a .

Example: $p = 2,\ g = 2$



good choice of a



bad choice of a

discriminant analysis

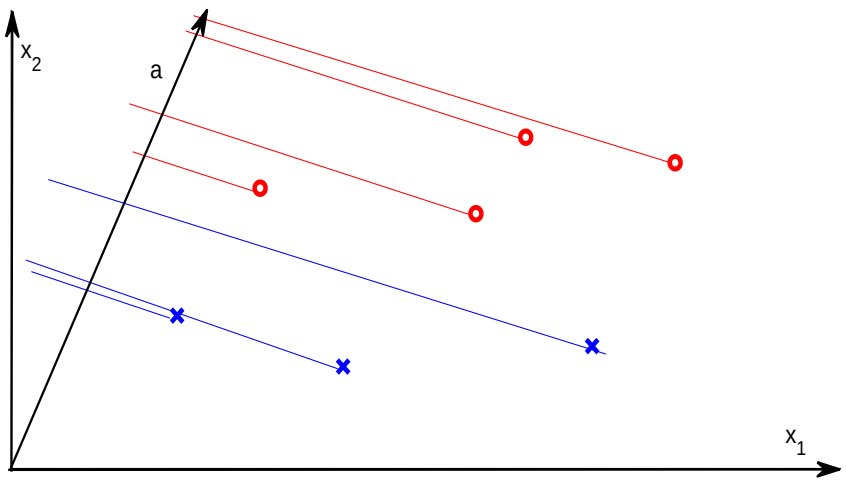
nonparametric ansatz by Fisher

$\mathbf{x} = (x_1, \ldots, x_p)^T$

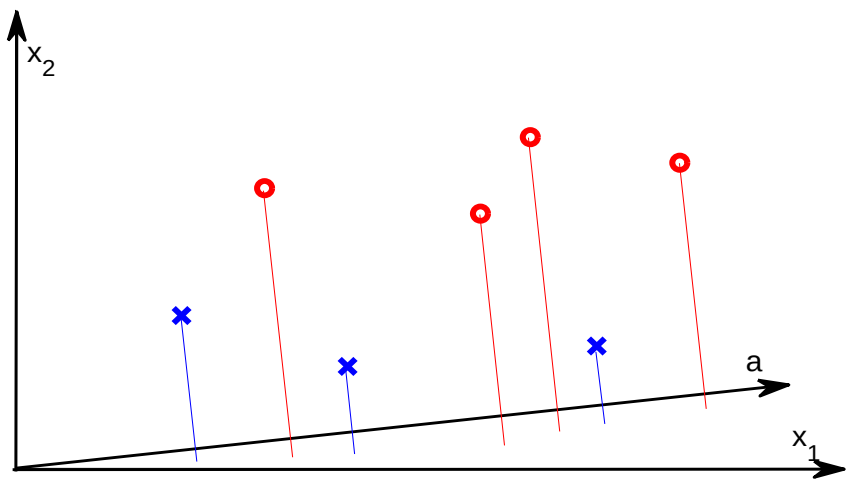
Idea: transform the p-dimensional problem to a one-dimensional problem.

How? With a linear combination of the vector $\mathbf{x}$:
$y = a^T\mathbf{x}\,,\ a^T = (a_1 \ldots a_p)$
i.e. $\mathbf{x}_{ki}\ , i = 1, \ldots, N_k$ is transformed to $y_{ki} = a^T\mathbf{x}_{ki}$.
With $\|a\| = 1$ the linear combination $a^T\mathbf{x}$ is the projection of the data $\mathbf{x}$ on a straight line with direction a .

Example: $p = 2,\ g = 2$



good choice of a



bad choice of a

discriminant analysis

nonparametric ansatz by Fisher

We have to choose a such that $Q(a) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_1^2 + S_2^2}$ with $\bar{y}_k = a^T \bar{\mathbf{x}}_k$ and $S_k^2 = \sum_{i=1}^{N_k} (y_{ki} - \bar{y}_k)^2$ is maximum (cf. anova). I.e. the variance between the groups should be maximum compared to the variance within the groups.

Different presentation of $S_1^2 + S_2^2$:

$$S_1^2 + S_2^2 = \sum_{i=1}^{N_1} a^T (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T a + \sum_{i=1}^{N_2} a^T (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)^T a =$$

$$= a^T W \cdot a \qquad W \ldots \text{ within group variance}$$

$$\Longrightarrow Q(a) = \frac{(a^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{a^T W \cdot a}$$

discriminant analysis

nonparametric ansatz by Fisher

We have to choose a such that $Q(a) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_1^2 + S_2^2}$ with $\bar{y}_k = a^T \bar{\mathbf{x}}_k$ and $S_k^2 = \sum_{i=1}^{N_k} (y_{ki} - \bar{y}_k)^2$ is maximum (cf. anova). I.e. the variance between the groups should be maximum compared to the variance within the groups.

Different presentation of $S_1^2 + S_2^2$:

$$S_1^2 + S_2^2 = \sum_{i=1}^{N_1} a^T (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T a + \sum_{i=1}^{N_2} a^T (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)^T a =$$

$$= a^T W \cdot a \qquad W \ldots \text{ within group variance}$$

$$\Longrightarrow Q(a) = \frac{(a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{a^T W \cdot a}$$

discriminant analysis

nonparametric ansatz by Fisher

We have to choose a such that $Q(a) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_1^2 + S_2^2}$ with $\bar{y}_k = a^T \bar{\mathbf{x}}_k$ and $S_k^2 = \sum_{i=1}^{N_k} (y_{ki} - \bar{y}_k)^2$ is maximum (cf. anova). I.e. the variance between the groups should be maximum compared to the variance within the groups.

Different presentation of $S_1^2 + S_2^2$:

$$S_1^2 + S_2^2 = \sum_{i=1}^{N_1} a^T (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T a + \sum_{i=1}^{N_2} a^T (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)^T a =$$

$$= a^T W \cdot a \qquad W \ldots \text{ within group variance}$$

$$\Longrightarrow \quad Q(a) = \frac{(a^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{a^T W \cdot a}$$

discriminant analysis

nonparametric ansatz by Fisher

$$\frac{\partial Q(a)}{\partial a} = \frac{2(a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)a^T W \cdot a - 2W \cdot a(a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{(a^T W \cdot a)^2} = 0$$

$$\Longrightarrow (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)a^T W \cdot a = W \cdot a \cdot a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = a \cdot \underbrace{\frac{a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{a^T W \cdot a}}_{\text{scalars, do not affect the direction of } a}$$

discriminant analysis

nonparametric ansatz by Fisher

$$\frac{\partial Q(a)}{\partial a} = \frac{2(a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)a^T W \cdot a - 2W \cdot a(a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2))^2}{(a^T W \cdot a)^2} = 0$$

$$\Longrightarrow (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)a^T W \cdot a = W \cdot a \cdot a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = a \cdot \underbrace{\frac{a^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{a^T W \cdot a}}_{\text{scalars, do not affect the direction of } a}$$

discriminant analysis

nonparametric ansatz by Fisher

Result: the linear Fisher discriminant function $y = a^T \mathbf{x}$ is for $g = 2$ identical to the discriminant function with assumed normal distribution with class-wise identical covariance matrices and ML decision rule (up to a constant term).

Fisher decision rule: Let $\mathbf{x}$ be an observation with unknown class index k. Compute $y = a^T \mathbf{x}$, the object is member of group 1 if y is closer to $\bar{y}_1$ than to $\bar{y}_2$:

$$\iff |y - \bar{y}_1| < |y - \bar{y}_2| \iff y > \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \iff$$

$$\iff a^T(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)) > 0$$

Conclusion: The linear discriminant analysis is relatively robust. The results are useful also if the assumption $\Sigma_k = \Sigma$ is violated.

discriminant analysis

nonparametric ansatz by Fisher

Result: the linear Fisher discriminant function $y = a^T \mathbf{x}$ is for $g = 2$ identical to the discriminant function with assumed normal distribution with class-wise identical covariance matrices and ML decision rule (up to a constant term).

Fisher decision rule: Let $\mathbf{x}$ be an observation with unknown class index k. Compute $y = a^T \mathbf{x}$, the object is member of group 1 if y is closer to $\bar{y}_1$ than to $\bar{y}_2$:

$$\iff |y - \bar{y}_1| < |y - \bar{y}_2| \iff y > \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \iff$$

$$\iff a^T(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)) > 0$$

Conclusion: The linear discriminant analysis is relatively robust. The results are useful also if the assumption $\Sigma_k = \Sigma$ is violated.

discriminant analysis

nonparametric ansatz by Fisher

Result: the linear Fisher discriminant function $y = a^T \mathbf{x}$ is for $g = 2$ identical to the discriminant function with assumed normal distribution with class-wise identical covariance matrices and ML decision rule (up to a constant term).

Fisher decision rule: Let $\mathbf{x}$ be an observation with unknown class index k. Compute $y = a^T \mathbf{x}$, the object is member of group 1 if y is closer to $\bar{y}_1$ than to $\bar{y}_2$:

$$\iff |y - \bar{y}_1| < |y - \bar{y}_2| \iff y > \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \iff$$

$$\iff a^T(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)) > 0$$

Conclusion: The linear discriminant analysis is relatively robust. The results are useful also if the assumption $\Sigma_k = \Sigma$ is violated.

discriminant analysis

nonparametric ansatz by Fisher

General case: g classes: We have the separation criterion:

$$Q(a) = \frac{\sum_{k=1}^{g} N_k(\bar{y}_k - \bar{y})^2}{\sum_{k=1}^{g} S_k^2} \longrightarrow \max!$$

We have $Q(a) = \frac{a^T B \cdot a}{a^T W \cdot a}$ with
$W = \sum_{k=1}^{g} W_k$, $W_k = \sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)^T$ and
$B = \sum_{k=1}^{g} N_k(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^T$. Let $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$ be the positive eigenvalues of $W^{-1}B$ ($q \leq \min\{p, g-1\}$) and $a_1 \ldots a_q$ the associated eigenvectors. Then we have:

$y_k = a_k^T \mathbf{x}$ reflects the given partitioning best for $k = 1$, second best for $k = 2$ etc.

discriminant analysis

nonparametric ansatz by Fisher

General case: g classes: We have the separation criterion:

$$Q(a) \;=\; \frac{\sum_{k=1}^{g} N_k(\bar{y}_k - \bar{y})^2}{\sum_{k=1}^{g} S_k^2} \;\longrightarrow\; \max!$$

We have $Q(a) = \frac{a^T B \cdot a}{a^T W \cdot a}$ with
$W = \sum_{k=1}^{g} W_k,\ W_k = \sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)^T$ and
$B = \sum_{k=1}^{g} N_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^T$. Let $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$ be the positive eigenvalues of $W^{-1}B$ ($q \leq \min\{p, g-1\}$) and $a_1 \ldots a_q$ the associated eigenvectors. Then we have:

$y_k = a_k^T \mathbf{x}$ reflects the given partitioning best for $k = 1$, second best for $k = 2$ etc.

discriminant analysis

nonparametric ansatz by Fisher

General case: g classes: We have the separation criterion:

$$Q(a) \;=\; \frac{\sum_{k=1}^{g} N_k(\bar{y}_k - \bar{y})^2}{\sum_{k=1}^{g} S_k^2} \;\longrightarrow\; \max!$$

We have $Q(a) = \frac{a^T B \cdot a}{a^T W \cdot a}$ with
$W = \sum_{k=1}^{g} W_k,\ W_k = \sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_i)^T$ and
$B = \sum_{k=1}^{g} N_k (\bar{\mathbf{x}}_k - \bar{\mathbf{x}})(\bar{\mathbf{x}}_k - \bar{\mathbf{x}})^T$. Let $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$ be the positive eigenvalues of $W^{-1}B$ ($q \leq \min\{p, g-1\}$) and $a_1 \ldots a_q$ the associated eigenvectors. Then we have:

$y_k = a_k^T \mathbf{x}$ reflects the given partitioning best for $k = 1$, second best for $k = 2$ etc.

discriminant analysis

nonparametric ansatz by Fisher

The y_k may be used all or just in part to reduce the dimension of $\mathbf{x}$:
$\mathbf{y} = (y_1 \dots y_r)^T = (a_1^T \mathbf{x} \dots a_r^T \mathbf{x})^T \quad r \leq q.$

We get the general Fisher decision rule: Choose $\hat{k}$ such that

$$\sum_{l=1}^{r} (a_l^T (\mathbf{x} - \bar{\mathbf{x}}_{\hat{k}}))^2 \leq \sum_{l=1}^{r} (a_l^T (\mathbf{x} - \bar{\mathbf{x}}_k))^2 \qquad \text{for} \qquad k = 1, \dots, g$$

where $r \leq q$.

For $p(1) = \dots = p(g)$ this rule is again equivalent to the ML decision rule with class-wise identical covariance matrices Σ (linear discriminant analysis).

discriminant analysis

nonparametric ansatz by Fisher

The y_k may be used all or just in part to reduce the dimension of $\mathbf{x}$:
$\mathbf{y} = (y_1 \dots y_r)^T = (a_1^T \mathbf{x} \dots a_r^T \mathbf{x})^T \quad r \leq q.$

We get the general Fisher decision rule: Choose $\hat{k}$ such that

$$\sum_{l=1}^{r} (a_l^T (\mathbf{x} - \bar{\mathbf{x}}_{\hat{k}}))^2 \leq \sum_{l=1}^{r} (a_l^T (\mathbf{x} - \bar{\mathbf{x}}_k))^2 \qquad \text{for} \qquad k = 1, \dots, g$$

where $r \leq q$.

For $p(1) = \dots = p(g)$ this rule is again equivalent to the ML decision rule with class-wise identical covariance matrices Σ (linear discriminant analysis).

discriminant analysis

nonparametric ansatz by Fisher

The y_k may be used all or just in part to reduce the dimension of $\mathbf{x}$:
$\mathbf{y} = (y_1 \dots y_r)^T = (a_1^T \mathbf{x} \dots a_r^T \mathbf{x})^T \quad r \leq q.$

We get the general Fisher decision rule: Choose $\hat{k}$ such that

$$\sum_{l=1}^{r} (a_l^T (\mathbf{x} - \bar{\mathbf{x}}_{\hat{k}}))^2 \leq \sum_{l=1}^{r} (a_l^T (\mathbf{x} - \bar{\mathbf{x}}_k))^2 \qquad \text{for} \qquad k = 1, \dots, g$$

where $r \leq q$.

For $p(1) = \dots = p(g)$ this rule is again equivalent to the ML decision rule with class-wise identical covariance matrices Σ (linear discriminant analysis).