7^{th} practice sheet multivariate methods II

Helmut Waldl

summer term 2012

Abbildung 1: Iris Setosa (Borsten-Schwertlilie), Iris Versicolor (Schillernde Schwertlilie) und Iris Virginica

25. Anderson measured 1935 the sepal length *ls*, the sepal width *bs*, the petal length *lp* and the petal width *bp* of the three species *Iris setosa*, *Iris versicolor* and *Iris virginica*:

ls	*bs*	*lp*	*bp*	spec	*ls*	*bs*	*lp*	*bp*	spec	*ls*	*bs*	*lp*	*bp*	spec	*ls*	*bs*	*lp*	*bp*	spec
50	33	14	02	set	64	28	56	22	vir	65	28	46	15	ver	67	31	56	24	vir
63	28	51	15	vir	46	34	14	03	set	69	31	51	23	vir	62	22	45	15	ver
59	32	48	18	ver	46	36	10	02	set	61	30	46	14	ver	60	27	51	16	ver
		⋮					⋮					⋮					⋮		

You find the data in the text file `irisdata.txt`. Use the SAS procedure CLUSTER to group the objects only using the variates *ls*, *bs*, *lp* and *bp*. Use single-, complete- and average-linkage-clustering with L_1- and L_2-distances (Hint: data-set type), as well as the centroid- and Ward's method and compare the results (no detailed interpretation of the SAS-output).

26. The L_q-distance between two p-dimensional vectors $\mathbf{x}_m$ and $\mathbf{x}_n$ is defined as follows:

$$d_q(\mathbf{x}_m, \mathbf{x}_n) := (\sum_{i=1}^{p} |x_{mi} - x_{ni}|^q)^{\frac{1}{q}}$$

Show that the L_q-distance is translation invariant but not scale invariant.

27. The Mahalanobis-distance between two p-dimensional vectors $\mathbf{x}_m$ and $\mathbf{x}_n$ is defined as follows:

$$d_M(\mathbf{x}_m, \mathbf{x}_n) := \sqrt{(\mathbf{x}_m - \mathbf{x}_n)^T K^{-1} (\mathbf{x}_m - \mathbf{x}_n)}$$

where $K = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ is the empirical covariance matrix. Show: d_M is invariant with respect to arbitrary non-singular linear transformations, , i.e. if C is an arbitrary regular $(p \times p)$-matrix and $b \in \mathbb{R}^p$ arbitrary and $\tilde{x}_n = C \cdot x + b$, then $d_M(\tilde{x}_n, \tilde{x}_m) = d_M(x_n, x_m)$

28. With Ward's method the homogeneity of the $(\nu - 1)$-th partition $\mathcal{C}^{\nu-1}$ is:

$$H(\mathcal{C}^{\nu-1}) = \sum_{k} \sum_{\mathbf{x}_n \in \mathbf{C_k}} \|\mathbf{x}_n - \bar{\mathbf{x}}_k\|^2$$

In the ν-th iteration the classes C_v and C_w with n_v and n_w objects are merged.

Show: The loss of homogeneity in the ν-th iteration is

$$H(\mathcal{C}^{\nu}) - H(\mathcal{C}^{\nu-1}) = \frac{n_v n_w}{n_v + n_w} \|\bar{\mathbf{x}}_v - \bar{\mathbf{x}}_w\|^2$$