Exercises for the course
Machine Learning 1
Winter semester 2016/17

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 5

Exercise 1: Finding the direction of maximal correlation between datasets (5+15+10 P)

In some applications, one might want to perform dimensionality reduction for two simultaneously acquired multivariate datasets. For example, in a neuroimaging experiment, one might simultaneously acquire functional magnetic resonance images (fMRI) and electroencephalographic (EEG) signals of the same participant. These two techniques measure quite different aspects of brain activity (fMRI measures blood flow while EEG measures electrical activity), and the measurements are generally differently scaled. Submitting the concatenated data to a joint PCA dimensionality reduction therefore bears the risk that the decomposition is dominated by one measurement modality at the expense of neglecting the other. On the other hand, performing separate PCAs bears the risk that the found principal subspaces miss interesting relationships between the two datasets that may be present in the original data.
The correlation coefficient

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

is a measure of the linear relationship between two variables that is independent of their scale and mean value. Assume the presence of two data sets $\mathcal{X} = \{\vec{x}_1, \dots, \vec{x}_N\}$ and $\mathcal{Y} = \{\vec{y}_1, \dots, \vec{y}_N\}$, where the samples $\vec{x}_i$ and $\vec{y}_i$ are measured at the same time, and where all $\vec{x}_i \in \mathbb{R}^{d_x}$ and all $\vec{y}_i \in \mathbb{R}^{d_y}$. Assume that all data have already been centered ($\sum_{i=1}^N \vec{x}_i = \vec{0}$, $\sum_{i=1}^N \vec{y}_i = \vec{0}$). We are interested in finding linear projections $\vec{w}_x \in \mathbb{R}^{d_x}$ and $\vec{w}_y \in \mathbb{R}^{d_y}$ that, when applied to $\mathcal{X}$ and $\mathcal{Y}$, maximize the correlation coefficient of the projected data.

(a) *Establish* the objective of this optimization problem as a function of $\vec{w}_x$ and $\vec{w}_y$.

(b) *Derive* analytic expressions for $\vec{w}_x$ and $\vec{w}_y$ at the optimum.

(c) *Derive* an analytic expression for the correlation coefficient at the optimum.

Note that characterizations in terms of (generalized) eigenvalue equations are sufficient. If several solutions exist, indicate which ones are the global optima.

Exercise 2: Fisher and Bayes (10+10 P)

In the asymptotic case where the sampled data for two classes tends to the probability density functions $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$ of means and covariances $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, and with class priors $P(\omega_1), P(\omega_2)$, Fisher linear discriminant finds a projection vector $\boldsymbol{w}^\star$ that maximizes the objective

$$J_{\mathrm{Fisher}}(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_B\, \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_W\, \boldsymbol{w}},$$

where $\boldsymbol{\Sigma}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top$ is the between-class covariance matrix and $\boldsymbol{\Sigma}_W = P(\omega_1)\,\boldsymbol{\Sigma}_1 + P(\omega_2)\,\boldsymbol{\Sigma}_2$ is the within-class covariance matrix. A solution to this problem is given in closed form as:

$$\boldsymbol{w}^\star = \boldsymbol{\Sigma}_W^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

The function $\phi(\boldsymbol{x}) = \langle \boldsymbol{w}^\star, \boldsymbol{x} \rangle$ is called the Fisher linear discriminant. Let ϕ be one element of the set of possible mappings Φ (e.g. given by the set of all possible one-dimensional linear projections $\mathbb{R}^d \to \mathbb{R}$ of the the data). We say that ϕ is optimal in the Bayes sense, if when building the decision boundary in the image of ϕ, no other mapping in the set Φ supports a decision boundary with lower expected error. For example, in the special case of two Gaussian distributions with covariances $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the Fisher linear discriminant is optimal in the Bayes sense when considering the set of all possible one-dimensional linear projections, and even when considering all possible functions.

(a) *Find* two *non*-Gaussian two-dimensional probability distributions $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$ with same covariances matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ such that the best one-dimensional linear projection in the Bayes sense is different from the best one-dimensional linear projection in the Fisher sense. Sketch these two distributions along with the best linear projection in the Fisher and Bayes sense.

(b) Consider now that the two classes are generated by two d-dimensional Gaussian distributions $p(\boldsymbol{x}|\omega_1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $p(\boldsymbol{x}|\omega_2) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. *Find* a mapping $\phi \in \mathbb{R}^d \to \mathbb{R}$ which for fixed mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ is optimal in the Bayes sense. *(Hint: your function ϕ should depend on $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, but* not *on the prior probabilities $P(\omega_1)$ and $P(\omega_2)$.)*

Exercise 3: Programming Exercise (50 P)

Download the programming files for exercise sheet 5 on ISIS and follow the instructions.