Machine Intelligence 1 **Exercise Sheet 7**
WS 2016/17, Obermayer/Böhmer due: 15.12.2016

Statistical learning theory

On the 8.12.2016 will be an irregular Deep-Learning tutorial by Youssef Kashef.
Homework of this exercise sheet is due on 15.12.2016!

Exercise T7.1: Empirical Risk Minimization (tutorial)

From the lecture you know that the number *linearly separable assignments* of p data points in N dimensions is

$$C_{(p,N)} = 2 \sum_{k=0}^{N-1} \begin{pmatrix} p-1 \\ k \end{pmatrix} .$$

(a) How is the Binomial coefficient $\binom{n}{k}$ defined?

(b) Use the theorem of binomial coefficients

$$\begin{pmatrix} r \\ q \end{pmatrix} + \begin{pmatrix} r \\ q-1 \end{pmatrix} = \begin{pmatrix} r+1 \\ q \end{pmatrix} \quad \text{to show that} \quad C_{(p,N)} + C_{(p,N-1)} = C_{(p+1,N)} .$$

(c) Explain how the number of linearly seperably assignments is related to the *VC-dimension* and over-fitting.

Exercise H7.1: Linear Discriminant Analysis (homework, 3 points)

A popular linear classifier known as *linear discriminant analysis* (LDA) can be motivated from conditional Gaussian density estimation for classes with equal covariances: assume that samples from class c are drawn according to the Gaussian density function

$$p(\underline{\mathbf{x}}|c) = \frac{1}{(2\pi)^{d/2}|\underline{\mathbf{\Sigma}}_c|^{1/2}} \exp\left(-\frac{1}{2}(\underline{\mathbf{x}} - \underline{\mu}_c)^\top \underline{\mathbf{\Sigma}}_c^{-1}(\underline{\mathbf{x}} - \underline{\mu}_c)\right) ,$$

where $\underline{\mu}_c \in \mathbb{R}^d$ is the *conditional mean* of class c, and $\mathbf{\Sigma}$ is the common covariance matrix. This allows to determine the conditional probability of the classes given the observed data and the classifier selects the class with the highest conditional probability.

(a) (2 points) For the case of 2 classes with equal covariances, i.e., $\underline{\mathbf{\Sigma}}_1 = \underline{\mathbf{\Sigma}}_2$, the decision boundary of the LDA classifier can be expressed as $\underline{\mathbf{w}}^\top \underline{\mathbf{x}} - b = 0$. Derive the parameters $\underline{\mathbf{w}} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ analytically.
Hint: the logarithm is monotonic, i.e. $a > b \Leftrightarrow \log(a) > \log(b)$.

(b) (1 point) Derive what shape of the boundary has when the covariance matrices are not equal.

Exercise H7.2: Variability of classification (homework, 4 points)

Assume data $x_\alpha \in \mathbb{R}^2$ drawn from two clusters, C_1 and C_2 and distributed according to the (multivariate) Normal distributions $\mathcal{N}(\underline{\mu}_i, 2\,\underline{\mathbf{I}})$, $i = 1, 2$ with $\underline{\mu}_1 = (0, 1)$ and $\underline{\mu}_2 = (1, 0)$. This task examines, how well a linear connectionist neuron can separate these two classes for increasing amounts N of available training data. Proceed as follows:

1. Generate a sample of $N/2$ data $\underline{\mathbf{x}}_\alpha$ from each of the two clusters. Let $y_\alpha = 1$ for $\underline{\mathbf{x}}_\alpha$ from C_1 and $y_\alpha = -1$ for $\underline{\mathbf{x}}_\alpha$ from C_2.
2. Find the weights of a linear connectionist neuron with output $y_\alpha = \underline{\mathbf{w}}^\top \underline{\mathbf{x}}_\alpha + b$ minimizing the squared error according to the analytical formula (e.g. Problem H5.2c in Ex. 5).
3. Find the predictions $\hat{y} = \mathrm{sign}(\underline{\mathbf{w}}^\top \underline{\mathbf{x}} + b)$ of this classifier for $N_{\mathrm{test}} = 1000$ new data drawn from the same distributions.
4. Calculate the percentage of correct classifications for the training (r_{train}) and test samples (r_{test}).

Repeat these steps 50 times for each $N \in \{2, 4, 6, 8, 10, 20, 40, 100\}$, and save the resulting parameters as well as the percentages for training and testing.

(a) (2 point) Plot the mean and standard deviation of r_{train}, and r_{test} against the number of samples N in an errorbar-plot (plotting N on the x-axis and the corresponding statistic on the y-axis).

(b) (1 point) Plot the means and standard deviation of w_1, w_2 and b against N.

(c) (1 point) Interpret your results! How do these estimates depend on N?

Exercise H7.3: The Binomial distribution (homework, 3 points)

This exercise examines the relation between the following 3 distributions:

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{(Binomial distribution)}$$

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{(Normal distribution)}$$

$$f(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{(Poisson distribution)}$$

(a) (1 point) Visualize the probability mass function $f(k; n, p)$ of the binomial distribution for a few different values of k, n, p. Describe an example random experiment, for which the binomial distribution might be a good model. What are the central properties of the binomial distribution and in which situations might it therefore not be a good model?

(b) (1 point) The normal distribution is sometimes used as an approximation to the binomial distribution. Under which conditions is this reasonable? Under which conditions is it problematic? Visualize one example where the Normal approximation is good and one where it is not. Give at least one reason why this distribution is so widely used. Is it a good approximation for the example random experiment you gave above?

(c) The Poisson distribution is often used as an alternative approximation to the binomial distribution. Under which conditions is it a good approximation? Visualize one example parametrization where the Poisson approximation is good and one where it is not. Is it a good approximation for the random experiment above?

Total 10 points.