# 2413, Machine Learning, Mock Exam
# University of Bern

20/12/2017

- **No books, notes, computers, calculators and cellular phones are allowed.**
- **This exam has 48 points in total.**
- **There are 6 questions.**

1. **[Total 20 points]** Give brief answers to the following questions.

   (a) **[2.5 points]** In the statements below, indicate whether they are TRUE or FALSE justifications regarding why test error could be less than training error.

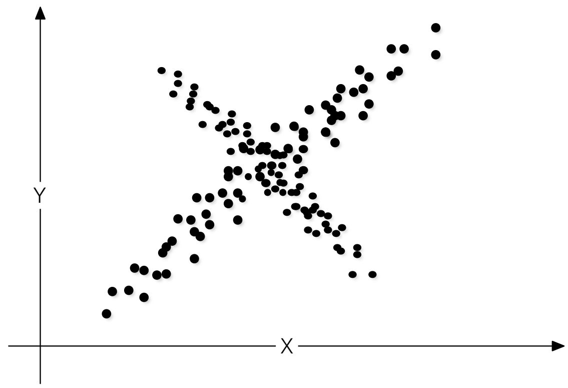
   - Test error is never less than training error. [TRUE/FALSE]
   - By chance the test set has easier cases than the training set. [TRUE/FALSE]
   - The model is too complex so training error overestimates test error. [TRUE/FALSE]
   - The model is too simple so training error overestimates test error. [TRUE/FALSE]

   **Solution:** 1.) False 2.) True 3.) False 4.) False

   (b) **[2.5 points]** It is a common practice in many machine learning algorithms to normalize the data, such that the data has zero mean and unit variance. If we normalize the data before applying k-means clustering, will we get the same cluster assignments as without normalization? Justify your answer.

   **Solution:** No, the issue is with the scaling applied when moving to unit variance. Centroid-assignments are computed according to euclidean distance and changing the scale of one of the variables can have an influence on this.

(c) **[2.5 points]** Suppose you are given the following set of points and run PCA. Draw the 1st and 2nd principal components in the figure below. Label them with $p_1$ and $p_2$.




**Solution:** First component along most variation... 2nd is orthogonal to it

(d) **[2.5 points]** The SVM problem is formulated as:

$$\begin{aligned} \hat{w}_C, \hat{b}_C, \hat{\boldsymbol{\xi}}_C = \arg \min_{w,b,\xi^{(i)}} \quad & \frac{1}{2}||w||^2 + C \sum_{i=1}^{m} \xi^{(i)} \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi^{(i)}, \quad i = 1, \ldots, m \\ & \xi^{(i)} \geq 0, \quad i = 1, \ldots, m \end{aligned} \tag{1}$$

Suppose we choose the parameter $C$ as follows:

i. Find the optimal parameters $\hat{w}_C, \hat{b}_C, \hat{\boldsymbol{\xi}}_C$ on the **training set** $\{x^{(i)}, y^{(i)}\}_{i=1,\ldots,m}$ for a range of values of $C = \{C_1, \ldots, C_K\}$.

ii. Evaluate the classification error $\hat{\epsilon}_{\text{test}}(\hat{w}_C, \hat{b}_C, \hat{\boldsymbol{\xi}}_C)$ of each optimal classifier $y = \hat{w}_C^\top x + \hat{b}_C$ on the **test set**. Choose the optimal $C^*$ as

$$C^* = \arg \min_C \hat{\epsilon}_{\text{test}}(\hat{w}_C, \hat{b}_C, \hat{\boldsymbol{\xi}}_C). \tag{2}$$

Is the classification error $\hat{\epsilon}_{\text{test}}(\hat{w}_{C^*}, \hat{b}_{C^*}, \hat{\boldsymbol{\xi}}_{C^*})$ a good estimate of the **generalization error**? Justify your answer.

**Solution:** No, by tuning the parameter on the test-set the model can be biased to particular features of the test-set (essentially overfit it in a sense). You should use a separate validation set for hyper-parameter tuning.

2. **[Total 5 points]** In linear regression we are given a training set with pairs $(\mathbf{x}^{(i)}, y^{(i)})$, $i = 1, \ldots, m$, and we look for a vector $\theta \in \mathbf{R}^n$ such that $y^{(i)} \approx \theta^T \mathbf{x}^{(i)}$.

   (a) **[2.5 points]** Describe what probabilistic assumptions lead to the maximum likelihood estimate of $\theta$.

   (b) **[2.5 points]** Show the distribution of $y^{(i)}$ given $x^{(i)}$ and parameterized by $\theta$ when the model error is Gaussian.

   **Solution**
   **[2.5 points]** Under the assumption that the noise of the data is IID. Also correct saying that in Least Square regression the noise is Gaussian.
   **[2.5 points]**

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big( - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\Big), \quad \epsilon^{(i)} \sim N(0, \sigma^2)$$

3. **[Total 5 points]** Write Jensen's inequality (when applied to expectations). What assumption needs to be satisfied for Jensen's inequality to be true?

   **Solution.** Jensen's inequality is $E[f(x)] \geq f(E[x])$ **(2p)** and it holds when $f$ is convex **(2p)**.

4. **[Total 8 points]** In the constrained optimization of $f$

$$\min_{\omega} f(\omega) \tag{3}$$
$$g_i(\omega) \leq 0 \qquad i = 1, \dots, m \tag{4}$$
$$h_j(\omega) = 0 \qquad j = 1, \dots, l \tag{5}$$

the corresponding generalized Lagrangian is

$$\mathcal{L}(\omega, \alpha, \beta) = f(\omega) + \sum_{i=1}^{m} \alpha_i g_i(\omega) + \sum_{i=1}^{l} \beta_i h_i(\omega) \tag{6}$$

where $\alpha_i > 0, \beta_i$ are Lagrange multipliers, $g_i$ are inequality constraints, and $h_i$ are equality constraints. There are four conditions for the Lagrange duality theory to guarantee that the primal and dual optimal solutions coincide in a convex optimization problem. One of them is the *complementary slackness condition*, $\alpha_i^* g_i(\omega^*) = 0$, $i = 1, \dots, m$.

(a) **[3 points]** List the other three conditions.

(b) **[5 points]** What are the effects of the complementary slackness condition on the optimal SVM classifier? Justify your answer.
**Hint**: The optimal SVM classifier can be written as

$$x^\top w^* + b^* = \sum_{i=1}^{m} \alpha_i^* y^{(i)} x^\top x^{(i)} + b^*$$

**Solution**
The KKT conditions are satisfied.
Such conditions are:

(a) **[1 points]** Primal feasibility $g_i(\omega^*) \leq 0$, $i = 1, \dots, m$ and $h_i(\omega^*) = 0$, $i = 1, \dots, p$;

(b) **[1 points]** dual feasibility $\alpha_i^* \geq 0$, $i = 1, \dots, m$;

(c) **[1 points]** Lagrangian stationarity $\nabla_x \mathcal{L}(\omega^*, \alpha^*, \beta^*) = 0$

(d) **[5 points]** From complementary condition, $\alpha_i > 0$ define the support vectors $x^{(i)}$. These are the only vectors left in the sum in the optimal classifier. Also, their function margin is exactly equal to one.

5. **[Total 10 points]** Consider the k-means objective for clustering data $\mathcal{X} = \{x^{(1)}, ..., x^{(N)}\}$ as $\mathcal{C} = \bigcup_{k=1}^{K} \mathcal{C}_k$ with cluster means $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$

$$\mathcal{J}(\mathcal{C}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \|x^{(i)} - \mu_k\|_2^2 \tag{7}$$

where $\mathcal{C}_k$ is a set that includes indices of the data points that belong to the $k$-th cluster and $\mathcal{C}_i \bigcap \mathcal{C}_j = \emptyset$ for all $i \neq j$ (so each point belongs to one cluster and only one).

Show that minimizing the above objective is equivalent to minimizing $\mathcal{J}'(\mathcal{C})$

$$\mathcal{J}'(\mathcal{C}) = \sum_{k=1}^{K} \frac{1}{2|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \|x^{(i)} - x^{(j)}\|_2^2 \tag{8}$$

**Hint**: Notice that $\mu_k = \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} x^{(i)}$.

**Solution.**

$$\mathcal{J}(\mathcal{C}) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} x^{(i)^T} x^{(i)} - 2x^{(i)^T} \mu_k + \mu_k^T \mu_k = \textbf{[3 points]} \tag{9}$$

$$\sum_{k=1}^{K} (\sum_{i \in \mathcal{C}_k} x^{(i)^T} x^{(i)} - \frac{2}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} x^{(i)^T} x^{(j)} + \frac{1}{|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} x^{(i)^T} x^{(j)}) = \textbf{[3.5 points]} \tag{10}$$

$$\sum_{k=1}^{K} \frac{1}{2|\mathcal{C}_k|} (2 \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} x^{(i)^T} x^{(i)} - 2 \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} x^{(i)^T} x^{(j)}) = \textbf{[3.5 points]} \tag{11}$$

$$\sum_{k=1}^{K} \frac{1}{2|\mathcal{C}_k|} \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \|x^{(i)} - x^{(j)}\|_2^2 \tag{12}$$

6. **[Total 10 points]** Assume that the samples $x^{(1)}, x^{(2)}, \dots, x^{(m)}$ are i.i.d. samples from a distribution described by the factor analysis model below,

$$z \sim \mathcal{N}(0, I), \tag{13}$$
$$\epsilon \sim \mathcal{N}(0, \Psi), \tag{14}$$
$$x = \mu + \Lambda z + \epsilon. \tag{15}$$

where $z$ and $\epsilon$ are independent.

**Hint**: Recall that $\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)\right]$.

(a) **[4 points]** Show that $x \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$.
**Solution.**
$E[X+Y] = E[X] + E[Y]$, $E[AX] = AE[X]$ and $E[X+a] = E[X] + a$ so $E(x) = \mu + \Lambda E(z) + E(\epsilon) = \mu$. **[1.5 point]**
$E[(x - E[x])(x - E[x])^T] = E[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T] = E[\Lambda z z^T \Lambda] + 2\Lambda E[z\epsilon^T] + E[\epsilon\epsilon^T] = \Lambda E[zz^T]\Lambda^T + \Psi$
note that $E[z\epsilon^T] = E[z]E[\epsilon^T] = 0$ because $z$ and $\epsilon$ are independent. **[2.5 points]**

**[6points]** What is the optimal $\mu$? Use the Maximum Likelihood estimation method, and maximise the log-likelihood.

**Solution. [3points]** The samples are drawn from the distribution $x \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$. The log-likelihood function according to the ML estimate is

$$\begin{aligned} l(\mu) &= \log \prod_{i=1}^{m} \frac{\exp(-\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu))}{(2\pi)^{n/2} |\Lambda\Lambda^T + \Psi|^{1/2}} \qquad (16) \\ &= \sum_{i=1}^{m} -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\Lambda\Lambda^T + \Psi|) + \\ &\quad \sum_{i=1}^{m} -\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) \end{aligned}$$

**[3points]** Note that the negative log-likelihood is a convex quadratic function in $\mu$, therefore we can find the optimal $\mu$ if we set the gradient to $0$. The gradient of the log-likelihood w.r.t. $\mu$ is

$$\begin{aligned} \nabla_\mu l(\mu) &= \nabla_\mu \sum_{i=1}^{m} -\frac{1}{2}(x^{(i)} - \mu)^T (\Lambda\Lambda^T + \Psi)^{-1} (x^{(i)} - \mu) \qquad (17) \\ &= \sum_{i=1}^{m} -(\Lambda\Lambda^T + \Psi)^{-1}\mu + (\Lambda\Lambda^T + \Psi)^{-1} x^{(i)}. \end{aligned}$$

From here, the solution is not very surprisingly,

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}. \tag{18}$$