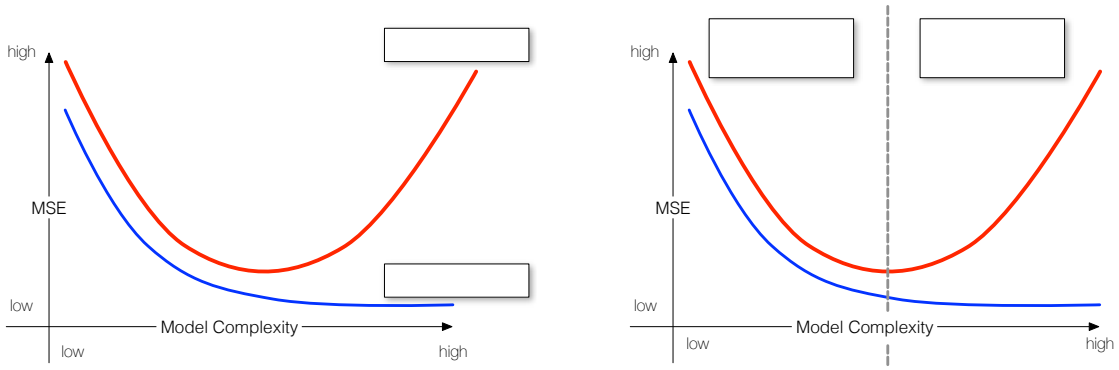
Machine Learning, Mock Exam 2
University of Bern

18/12/2019

- **No books, notes, computers, calculators and cellular phones are allowed.**
- **This exam has 23 points in total.**
- **There are 4 questions.**

1. **[Total 10 points]** Give brief answers to the following questions.

   (a) **[2 points]** The two graphs below show train and test-error of some model as a function of the model complexity. In the left graph, indicate which of the two curves shows the 'training error' or 'test error'. In the right graph, indicate which regions show 'high variance' or 'high bias'.



   Solution: 1.) Upper curve is test-error, lower training error. 2.) Left part is high bias, right is high variance

   (b) **[2 points]** In the statements below, indicate whether they are TRUE or FALSE justifications regarding why test error could be less than training error.

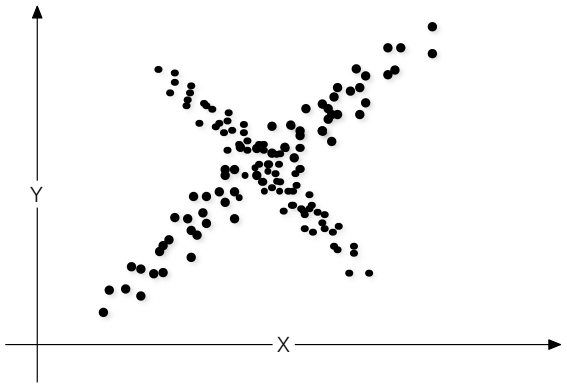
   - Test error is never less than training error. [TRUE/FALSE]
   - By chance the test set has easier cases than the training set. [TRUE/FALSE]
   - The model is too complex so training error overestimates test error. [TRUE/FALSE]
   - The model is too simple so training error overestimates test error.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

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

   (c) **[2 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.

(d) **[2 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

(e) **[2 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 4 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 $E[f(x)] \geq f(E[x])$ **[2 points]** and it holds when f is convex **[2 points]**.

3. **[Total 5 points]** Assume that the samples $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ are i.i.d. samples from a distribution described by the factor analysis model below,

$$z \sim \mathcal{N}(0, I), \tag{3}$$
$$\epsilon \sim \mathcal{N}(0, \Psi), \tag{4}$$
$$x = \mu + \Lambda z + \epsilon. \tag{5}$$

where z and ϵ are independent.

Show that $x \sim \mathcal{N}(\mu, \Lambda\Lambda^T + \Psi)$.

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

Solution:

Because x is a linear combination of Gaussian random variables, we know that it will also be Gaussian. It suffices to compute the mean and covariance of this Gaussian.

We have $E[x] = \mu + \Lambda E[z] + E[\epsilon] = \mu$, because $E[X+Y] = E[X]+E[Y]$, $E[AX] = AE[X]$ and $E[X + a] = E[X] + a$. **[2 points]**

We also have $E[(x - E[x])(x - E[x])^T] = E[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T] = E[\Lambda z z^T \Lambda] + 2\Lambda E[z\epsilon^T] + E[\epsilon\epsilon^T] = \Lambda E[zz^T]\Lambda^T + \Psi$.

Note that $E[z\epsilon^T] = E[z]E[\epsilon^T] = 0$ because z and ϵ are independent. **[3 points]**

4. **[Total 4 points]** Suppose you are given a set of training samples $\left\{x^{(i)}; i = 1, \ldots, m\right\}$, where $x^{(i)} \in \mathbb{R}^n$. You want to use Principal Component Analysis (PCA) to represent this data as $\left\{y^{(i)}; i = 1, \ldots, m\right\}$ in $k < n$ dimensions (*i.e.*, $y^{(i)} \in \mathbb{R}^k$).

   (a) **[Total 2 points]** What is the functional relationship between the latent representation $y^{(i)}$ and the observed data $x^{(i)}$? Clearly identify the parameters which are learned.

   Solution: The relationship is

$$y^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ \vdots \\ u_k^T x^{(i)} \end{bmatrix}$$

   And the learned parameters are $\{u_i; i = 1, \ldots, k\}$.

   (b) **[Total 2 points]** Consider the case of $k = 1$. Specify the objective function (including constraints) that is optimized when performing PCA (after the normalization step).

   Solution:

$$\begin{aligned} \max_u \quad & \frac{1}{m} \sum_{i=1}^{m} u^T x^{(i)} x^{(i)T} u \\ \text{s.t.} \quad & \|u\|_2 = 1 \end{aligned}$$