T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

18 February 2014.

To pass the course you must also pass the term project. Results of this examination are valid for one year after the examination date.

This examination has five problems each worth 10 points, and two pages. You must answer in English. Please write clearly and leave a wide left or right margin. You can have a calculator, with memory erased. No other extra material is allowed.

The results will be announced in Noppa. You can keep this paper.

Problem 1: Explanations of concepts (10 points total)

Explain the terms below in the context of the course. If two terms are given, explain them so that is becomes clear what they have in common and what are the differences. Use full sentences.

1. Akaike Information Criterion (AIC)—Bayesian Information Criterion (BIC) (2 points)
2. Forward search, in context of feature selection (2 points)
3. parametric methods—nonparametric methods (2 points)
4. Receiver operating characteristics (ROC) curve (2 points)
5. Bayes estimate - Maximum a posteriori (MAP) estimate (2 points)

Problem 2: Model Selection (10 points total)

- Explain what is the K-fold Cross Validation Scheme and write down one advantage and disadvantage of it. (2 points)
- Consider a parametric regression (Bayesian regression) scenario where we try to regress target values r from one-dimensional inputs x using the regressor $r \approx g(x|\theta) = a_0 + a_1 x + a_2 x^2$, where the parameters are $\theta = (a_0, a_1, a_2)$. We assume Gaussian noise $p(r|x, \theta) \sim N(g(x|\theta), \sigma^2)$ where σ^2 is the noise variance which is a known constant.

  Assume we have observed N input points x_i and their corresponding targets r_i. Write down the likelihood function of the model. (3 points)
- Now suppose that we will also assume a Gaussian prior for the parameters, $a_0 \sim N(0, 1/\lambda)$, $a_1 \sim N(0, 1/\lambda)$, $a_2 \sim N(0, 1/\lambda)$, where $1/\lambda$ is the prior variance of the parameters. Write down the equation you need to maximize in order to find the maximum a posteriori (MAP) estimate of the parameters. (3 points)
- Explain the role of the constant λ in the prior distribution, and how it causes regularization of the learned parameters θ. (2 points)

Problem 3: Clustering (10 points total)

a) Define the difference between hard clustering and fuzzy clustering. Name an example method for both types of clustering. (2 points)

b) Write the pseudocode for EM algorithm for Clustering. Define all the terminologies properly. (4 points)

c) You have observed a data set of five one-dimensional points whose coordinates are $\{0, 1, 3, 4, 5\}$. You want to cluster the set of points into two clusters using the K-means algorithm ($K = 2$). Initialize the centroids of the two clusters at locations 2.5 and 4.5 respectively. Run the K-means algorithm for two iterations. At the start of the algorithm and after each of the two iterations, describe the clusters (which points belong to each cluster, what are the locations of the cluster centroids). (4 points)

Problem 4: Principal Component Analysis (10 points total)

You have a data set of the following five two-dimensional points:

$$\mathcal{X} = \left\{ \begin{bmatrix} -5 \\ 4 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -3 \end{bmatrix}, \begin{bmatrix} 5 \\ -4 \end{bmatrix} \right\}$$

You want to reduce the dimensionality of the data points to one, using Principal Component Analysis. You have already estimated that the data is zero-mean and has a covariance matrix of

$$\mathbf{S} = \begin{bmatrix} 10 & -8 \\ -8 & 10 \end{bmatrix}$$

and you know the covariance matrix can be diagonalized as $\mathbf{C}^\top \mathbf{S} \mathbf{C} = \mathbf{D}$ where

$$\mathbf{C} = \begin{bmatrix} -1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 2 & 0 \\ 0 & 18 \end{bmatrix}.$$

Perform the following tasks.

a) Explain how the matrices C and D are related to Principal Component Analysis. (1 points)

b) Reduce the dimensionality of the data to one, by computing the projections of the five data points onto the first principal component. It is enough to do the computation for the first two data points. (3 points)

c) Compute the proportion of variance explained by the first principal component. (2 points)

d) Reconstruct the original data points approximately, by projecting the coordinates computed in step a) back into the original space. It is enough to do the computation for the first two data points. (2 points)

e) Compute the reconstruction error. If you reconstructed just the first two data points in step c), it is acceptable to use only those two points in this step. (2 points)

Problem 5: Nonparametric Density Estimation (10 points total)

a) In the context of probability density estimation, explain the naive estimator and give its mathematical definition. (2 points)

b) In the context of density estimation, explain the kernel estimator and give its mathematical definition. (2 points)

c) You have observed eight one-dimensional data points whose coordinates are shown in the figure below. Use the naive estimator with bin width $h = 4$ to estimate the probability density over the interval $[-6, 6]$. Draw the density function. (3 points)

d) Using the same naive estimator as in part c), give the numerical value of the probability density at the locations -4.5, 0.5, and 1.75. (3 points)

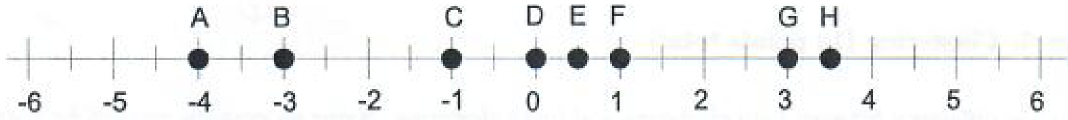

Figure 1: Observed data points for problem 5 d). For convenience the data points are labeled A-H.