T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

25 October 2013.

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. consistent hypothesis—version space (2 points)
2. overfitting—underfitting (2 points)
3. histogram estimator—naive estimator, in context of density estimation (2 points)
4. expected utility in classification (2 points)
5. k-fold cross-validation (2 points)

Problem 2: Bayesian Decision Theory and Parametric Methods (10 points).

a) Consider $X_1, X_2, \ldots, X_n$ are i.i.d. observations from a model $P(X|\theta)$ with unknown parameter θ. If you want to estimate θ following Bayes Theorem, answer the following first:

- What do the concepts prior, likelihood and posterior mean in the above problem? Write with mathematical notation. You do not have to define specific functions for the prior, likelihood, and posterior, just explain what the concepts are. (2 points)
- How can you compute the posterior if you know the prior and likelihood? (2 points)
- If you know the posterior density of θ, how can you compute the Bayes estimate of θ? You do not have to perform the computation, just explain how it would be done. (2 points)

b) Suppose in a), the observations $X_1, X_2, \ldots, X_n$ are light bulbs where each bulb X_i is either working ($X_i = 0$) or broken ($X_i = 1$), and we have observed $n = 10000$ bulbs. We assume the model $P(X|\theta)$ is a Bernoulli process, where the parameter θ, $0 < \theta < 1$, is the probability for a bulb being broken. Then answer the following:

- We want to use a flat prior (also known as a uniform prior) for θ. Write the equation of the prior. (1 point)
- Write the expression of the posterior density of θ. (3 points)

Problem 3: Clustering (10 points).

- Give one example of Hard and Fuzzy Clustering (also called Soft Clustering). Explain the differences between these two types of clustering. (2 points)
- Suppose you are performing an iteration of K-means clustering, and you know the set of K cluster means m_i. What would be the error function that you need to minimize to assign observations to the clusters? (2 points)
- Write the Lloyd's algorithm in pseudocode. (4 points)
- Does the solution of K-means depend on the initial location of the cluster means? If yes, how can you try to get better solutions? If not, why not? (2 points)

Problem 4: Principal Component Analysis (10 points total)

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

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

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

$$\mathbf{S} = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

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

$$\mathbf{C} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \text{ and } \mathbf{D} = \begin{bmatrix} 16 & 0 \\ 0 & 4 \end{bmatrix}.$$

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 Classification (10 points).

You have acquired the training data shown in the scatter plot below, where circles are locations of data points, '+' signs are data from the positive class and '-' signs are data from the negative class. You also have three validation points marked as 1, 2, and 3 in the scatter plot. You know that validation point 1 comes from the positive class and validation points 2 and 3 come from the negative class.

a) Explain the principle of k-nearest neighbor classification. Write the necessary equations for the case $k = 1$. (3 points)
b) Classify the three validation points based on the training set, using k-nearest neighbor classification with $k = 1$. (2 points)
c) Classify the three validation points based on the training set, using k-nearest neighbor classification with $k = 5$. (3 points)
d) Compute the classification errors on the validation set, and choose the best complexity for the classifier (choose $k = 1$ or $k = 5$). (2 points)