T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES

EXAMINATION 24 October 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 three 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: Multiple-choice questions (10 points total).

The following five questions each have different proposed answers, in each question only one answer is correct. For each question, give both your answer (one of A,B,C,...) and your confidence ("High" or "Low"). For example, "A, Low" is a proper way to answer a question. Each question is graded as follows:

- +2 if the answer is correct and confidence is High
- +1 if the answer is correct and confidence is Low
- 0 if the answer is missing
- −1 if the answer is wrong and confidence is Low
- −2 if the answer is wrong and confidence is High.

The total score for the multiple-choice questions is between 0 and 10 (you cannot get a negative total score).

1) Assume the covariance matrix of your dataset $\mathcal{X}$ is Σ and while doing Principal Component Analysis, you found $\Sigma = \mathbf{CDC}^T$. Here $\mathbf{D}$ is the following diagonal matrix:

$$\begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2.1 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & .9 \end{pmatrix}$$

Then the proportion of variance (POV) explained by the first 2 principal components would be (accurate until 2 decimal places):

A) .85
B) .68
C) .31
D) 1
E) None of the above answers is correct.

2) For a dataset distributed as i.i.d $\mathcal{N}(\mu, \sigma)$, after computing the posterior distribution using a normal prior for μ you found the posterior distribution of μ to be $\mathcal{N}(5.24, 4.62)$. Then the MAP-estimate and Bayes-estimate of μ would be:

A) MAP-estimate = 5, Bayes-estimate = 4;
B) MAP-estimate = 2.62, Bayes-estimate = 2.62;
C) MAP-estimate = 5.24, Bayes-estimate = 5.24;
D) MAP-estimate = 5.24, Bayes-estimate = 2.62;
E) None of the above answers is correct.

3) EM-algorithm is a type of fuzzy clustering as:

A) The clusters we get are fuzzy-sets;
B) Each observations are assigned to only one of the K-clusters;

C) For each observation there are different probabilities of it being assigned to K different clusters;

D) The algorithm uses fuzzy-methods;

E) None of the above answers is correct.

4) Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be written in a general form as $\mathcal{L}_{ML}(\theta \mid \mathcal{X})$-regularizer term. The regularizer term of BIC can be expressed as:

A) $\frac{1}{2}N\times$(regularizer term of AIC);

B) $\frac{1}{2}logN\times$(regularizer term of AIC);

C) $\frac{1}{N}\times$(regularizer term of AIC);

D) $\frac{1}{N}N\times$(regularizer term of AIC).

5) In a classification problem, you have two classes 0 and 1. You observed a new data point x_{new}, whose likelihood of belonging to Class 0 and Class 1 are correspondingly .32 and .41. Your prior beliefs (probabilities) for the Class 0 and Class 1 are correspondingly 0.6 and 0.4. Then, following Bayesian decision theory, x_{new} belongs to:

A) Not enough information given to use Bayesian Decision Theory;

B) Class 0;

C) Class 1;

D) It can belong to both Class 0 and 1 with equal probability.

Problem 2: 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 it becomes clear what they have in common and what are the differences. Use full sentences.

1. Forward Search, in context of feature selection (2 points)

2. naive estimator—kernel estimator, in context of density estimation (2 points)

3. regression—classification (2 points)

4. Bias/Variance Dilemma (2 points)

5. parametric methods—nonparametric methods (2 points).

Problem 3: Bayesian Decision Theory and Parametric Methods (10 points total).

a) A company has to decide whether to accept or reject a lot of incoming parts. (Label these actions a_1 and a_2 respectively.) The lots are of three types: θ_1 (very good), θ_2 (acceptable), and θ_3 (bad). The loss $L(\theta_i, a_j)$ incurred in making the decision is given in the following table. The prior belief is that $\pi(\theta_1) = \pi(\theta_2) = \pi(\theta_3) = \frac{1}{3}$.

	a_1	a_2
θ_1	0	3
θ_2	1	2
θ_3	3	0

- What is the definition of expected utility - explain with equations? (1 points)
- What is the expected utility of action a_1 for the company? (2 points)
- Find the optimal decision for the company. (2 points)

b) A smart-phone producing company periodically samples smart-phones coming off a production line, in order to make sure the production process is running smoothly. They choose a sample of size 5 and observe the number of defectives (X). They assume number of defectives (X) is distributed as $Binomial(5, \theta)$. Past records show that the proportion of defectives θ varies according to a $Beta(1, 9)$ distribution.

[Hints: (a) $Binomial(x|n, \theta) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{(n-x)}$; (b) $Beta(\theta|\alpha, \beta) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha,\beta)}$; (c) $B(\alpha, \beta) = \int_0^1 x^{(\alpha-1)}(1-x)^{(\beta-1)}$]

- If they have observed $X = 2$ defective smart-phones in one inspection run, explain how we derive the posterior distribution of θ? (3 points) [Hint: The posterior distribution would be $Beta(\alpha + 2, \beta + (5 - 2))$]
- What would be the Bayes estimate of θ using the posterior distribution dervied in last step? [Hint: You know the mean of $Beta(\alpha, \beta)$ distribution is $\frac{\alpha}{\alpha+\beta}$] (2 points).

Problem 4: Clustering (10 points total).

- What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis? (3 points)
- In EM-algorithm if the probability of assignment of the i-th observation to one of the K clusters ($\{h_i^k\}_{k=1,\ldots,K}$) becomes 1 for all observations (e.g. for $i = 8, h_8^3 = 1$, for $i = 5, h_5^2 = 1$ and so on), then show that the EM-algorithm becomes Lloyd's algorithm for K-means clustering. (3 points)
- Do three iterations of Lloyd's algorithm for K-means clustering on the 2-dimensional data below. Use $K = 2$ clusters and the initial prototype vectors (=mean vectors) $m_1 = (0.0, 2.0)$ and $m_2 = (2.0, 0.0)$. Write down calculation procedure and the cluster memberships as well as mean vectors after each iteration. Draw the data points, cluster means and cluster boundary after each iteration.

t	x^t
1	(0.0, 1.0)
2	(1.0, 2.0)
3	(4.0. 5.0)
4	(5.0, 3.0)
5	(5.0, 4.0)

(4 points).

Problem 5: Principal Component Analysis (10 points total).

You have a dataset of N two-dimensional points $\mathbf{y}^t$. You want to perform Principal Component Analysis (PCA) on the dataset. You have already estimated that the data is zero-mean and has the covariance matrix

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

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

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

Perform the following tasks:

a) Explain how the matrices $\mathbf{C}$ and $\mathbf{D}$ are related to Principal Component Analysis. (2 points)
b) In 2-dimensional space, plot the PCA coordinates (directions of the largest and second-largest variance of the data) (3 points)
c) Compute the proportion of variance explained by the first principal component. (2 points)
d) Define the principal components of the data by $\mathbf{z}^t = \mathbf{C}^T\mathbf{y}^t$. What is the covariance matrix of $\mathbf{z}^t$? (3 points).