T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

26 October 2012.

To pass the course you must also submit 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 4 pages. You have to 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 on November 25, at latest.

You can keep this paper.

1. Multiple choices questions (10 points). The following five questions have different proposed answers. Only one of them is correct. You have to give your answer along with your confidence ("High" or "Low") for each answer. Grading for each of these questions is then:

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

   Write on your answer sheet the correct answer A, B, C, D,...) along with the confidence you have (High or Low) for that question. For example, "A, Low" is a proper way of answering a question. No need to justify your answers. Total score for this question is between 0 and 10 (you cannot get a negative score for the whole question).

   1) For a binary classification problem, each class is modeled using a Multivariate Normal (Gaussian) Distribution. A Bayes classifier is calculated.

      A) The boundary is always linear.
      B) The boundary is always nonlinear.
      C) The boundary is independent from the priors of the classes.
      D) The boundary can never separate the classes perfectly (for the training set).
      E) None of the previous answers is correct

   2) For a multidimensional dataset, a Principal Component Analysis (PCA) is performed.

      A) The average reconstruction error is never increasing with the dimension of projection. / high
      B) The projection is independent from the variances of the input variables.

C) The average reconstruction error is always increasing with the dimension of projection.

D) The projection dimension has to be larger than the number of points and the number of variables (samples).

E) None of the previous answers is correct

3) The Lloyd's algorithm is used to perform clustering.

A) This algorithm will never converge and has to be stopped after an arbitrary number of iterations.

B) The error function which is minimized can increase for some iterations but is globally decreasing.

C) The Lloyd's algorithm will always converge to the best clustering solution.

D) The Lloyd's algorithm is independent from the initialization.

E) None of the previous answers is correct

4) For a binary classification problem, a K-Nearest-Neighbor (KNN) Classifier is built.

A) The classification error is always decreasing with respect to the parameter K.

B) The best value for K is always 1.

C) The parameter K can be optimized using validation.

D) The performances of the KNN classifier are independent from the distance metric which is used.

E) None of the previous answers is correct

5) A k-fold cross-validation is used to determine the optimal complexity of a regression model.

A) The cross-validation error is a perfect estimate of the generalization performances of the regression model.

B) The best value for k is always 2.

C) The best value for k is always 10.

D) The complexity selected by the k-fold cross-validation is always larger than the complexity selected using a Bayesian Information Criterion (BIC) regularization.

E) None of the previous answers is correct

2. *Model selection.* Assume that you have at your disposal a data set $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t=1}^N$, where r^t is a class and $\mathbf{x}^t$ is a covariate; and a set of k black box classification algorithms A_i, $i \in \{1, \dots, k\}$, which try to predict the class r, given the covariate $\mathbf{x}$ and the training data. More formally, you can think A_i as a known arbitrary function $r_{PREDICTED} = A_i(\mathbf{x}, \mathcal{X}_{TRAIN})$, where $r_{PREDICTED}$ is the predicted class, given $\mathbf{x}$, and $\mathcal{X}_{TRAIN}$ is the data used to train the classifier. Your task is to choose and train the classification algorithm that would work best for yet unseen data. Describe, in detail, different ways how you could accomplish this (and why). How do you expect the various classification errors to behave? (10 points)

3. (a) *Maximum Likelihood* (4 points). Consider a univariate data set $\mathcal{X} = (x^1, x^2, \ldots, x^N)$ that has a *log-normal* distribution. Find the maximum likelihood estimates of the mean μ and variance σ^2. The probability density function is given by

$$p(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0.$$

(b) *Naïve Bayes* (6 points). Consider binary classification for multivariate data $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t\in\{1,\ldots,N\}}$, where $r^t \in \{0, 1\}$ and $\mathbf{x}^t \in \mathbb{R}^d$. Assume that

- r is Bernoulli distributed with $P(r = 1) = \pi$.
- Variable x_i, $i = 1, \ldots, d$ is continuous and normally distributed with $P(x_i|r = k) = \mathcal{N}(\mu_{ik}, \sigma_i^2)$. The variance σ_i^2 is class independent!
- All variables are independent of each other given the class label r (Naïve Bayes assumption).

Show that the posterior distribution $P(r = 1|\mathbf{x})$ can be written in logistic form, i.e.

Bayes $P(r=1|x) = \frac{P(r=1)P(x|r)}{[illegible]}$

$P(r=1)$ [illegible]

$$P(r = 1|\mathbf{x}) = \frac{1}{1 + \exp(w_0 + \sum_{j=1}^{d} w_j x_j)}.$$

and write down the expressions for w_0 and w_j, $j = 1, \ldots, d$.

4. *Feature selection.* Consider the feature selection in classification problems.

(a) What is feature selection and why is it needed? (4 points)

(b) Assume that you have a binary classification algorithm. Explain, also using pseudocode, how you would implement forward and backward selection of features (in a real world application). (4 points)

(c) What can you say about time complexity and the optimality of the solutions produced by the forward and backward selection methods? (2 points)

5. *Combining classifiers* (a) Explain why is it a good idea to teach several different classifiers and use majority voting as the final classification. (2 points) (b) Why does this approach work better if the individual base-learners are as different as possible? (2 points) (c) Give at least four ways to make them different. (4 points) (d) Assuming each base learner gives a correct classification with probability p and the classification errors are independent of each other, what is the probability that a majority vote over L classifiers gives the correct answer? (2 points)