



KTH Computer Science
and Communication

Exam in DD2431 Machine Learning

2015-10-27, kl 14.00 – 18.00

Aids allowed: *calculator, language dictionary.*

A Questions for pass or fail

Only one alternative is correct for each question.

Note: To pass the exam you must give the correct answer on almost *all* questions in this section. Only *one* error will be accepted, so be very careful not to make any unnecessary mistakes here.

A-1 Probabilistic Learning

The log likelihood of a data set composed by independent and identically distributed observations can be computed

- a) as the product of the log likelihoods of each individual observation.
- b) as the sum of the log likelihoods of each individual observation.
- c) as the integral of the probability distribution function.
- d) using a conjugate prior.

A-2 Decision Trees

What principle is commonly used when building a decision tree?

- a) Maximize the number of leaf nodes.
- b) Maximize the projection length.
- c) Maximize the distance to borderline samples.
- d) Choose questions that maximize information gain.

A-3 Overfitting

You have trained a model (classifier) using some training sample data. Under which conditions is overfitting most likely to occur?

- a) Few training samples are used to train a relatively simple model.
- b) Many training samples are used to train a relatively simple model.
- c) Few training samples are used to train a relatively complex model.
- d) Many training samples are used to train a relatively complex model.

A-4 Kernel function

What is the role of the *kernel function* in a Support Vector Machine?

- a) Perform scalar products in a virtual high-dimensional space.
- b) Low-pass filter the input to reduce the effect of noise.
- c) Increase the length of the weight vector.
- d) Separate positive from negative samples.

A-5 Artificial Neural Networks

What is the main reason for using *hidden layers* in a feed-forward artificial neural network?

- a) The network will generalize better.
- b) More training data can be processed in parallel.
- c) Non-linear decision boundaries can be learned.
- d) Noisy data will be filtered away before reaching the classifier.

A-6 Regression and Classification

Which one of the following is true as the main difference between regression and classification?

- a) Techniques of classification require sampling while those of regression require pruning.
- b) The output is discrete for classification and continuous for regression.
- c) Nearest neighbours can be used for regression but not for classification.
- d) Classification is based on supervised learning whereas regression is based on unsupervised learning.

A-7 Ensemble Learning

Weak learners are trained and combined in different Ensemble methods in different manners. Which one of the following best describes the characteristics of the training process?

- a) In Bagging and Decision Forests, the weak learners are trained in a parallel manner.
- b) In Boosting and Bagging, the weak learners are trained in a parallel manner.
- c) In Decision Forests and Boosting, the weak learners are trained in a parallel manner.
- d) In Bagging, Boosting, and Decision Forests, the weak learners need be trained sequentially.

A-8 Principal Component Analysis (PCA)

Which one of the following statements is *incorrect*?

- a) PCA serves for subspace methods to represent the data distribution in each class.
- b) PCA is useful for reducing the effective dimensionality of data.
- c) PCA can be seen as a tool for acquiring an approximated representation of data.
- d) PCA is a supervised learning method that requires labeled data.

B Questions for higher grades

Preliminary number of points required for different grades:

$$22 \leq p \leq 24 \rightarrow A$$

$$18 \leq p < 22 \rightarrow B$$

$$12 \leq p < 18 \rightarrow C$$

$$6 \leq p < 12 \rightarrow D$$

$$0 \leq p < 6 \rightarrow E$$

B-1 Terminology

(4p)

For each term (a–h) in the left list, find the explanation from the right list which best describes how the term is used in machine learning.

- | | |
|-----------------------------|--|
| | 1) Location of concentration of a set of data |
| | 2) Samples on the margin of the decision surface |
| | 3) An approach to iteratively fitting model parameters with latent variables |
| a) k -nearest neighbour | 4) A situation with short of samples in the training phase |
| b) Support Vector | 5) A learning approach by ensemble machines |
| c) Perceptron learning | 6) Measure of spread of a random variable |
| d) Categorical distribution | 7) A subportion of area defined by two sets of parallel lines |
| e) Subspace | 8) Evidence for a specific hypothesis |
| f) Dropout | 9) A space spanned by a set of linearly independent vectors |
| g) EM-algorithm | 10) Probability divided by the bias |
| h) Variance | 11) Distribution of discrete stochastic variables |
| | 12) A method of regularization used in deep neural networks |
| | 13) Class prediction by a majority vote |
| | 14) Error driven method to compute weights in a single layer network |
| | 15) Learning based on distorted versions of the training data |

B-2 Classification

(2p)

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use Naïve Bayes classifier and get an error rate of 20% on the training data. We also get the average error rate (averaged over both test and training data sets) of 24%. Next we use 1-nearest neighbor and get an average error rate (averaged over both test and training data sets) of 16%.

- What was the error rate with Naïve Bayes classifier on the test set?
- Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

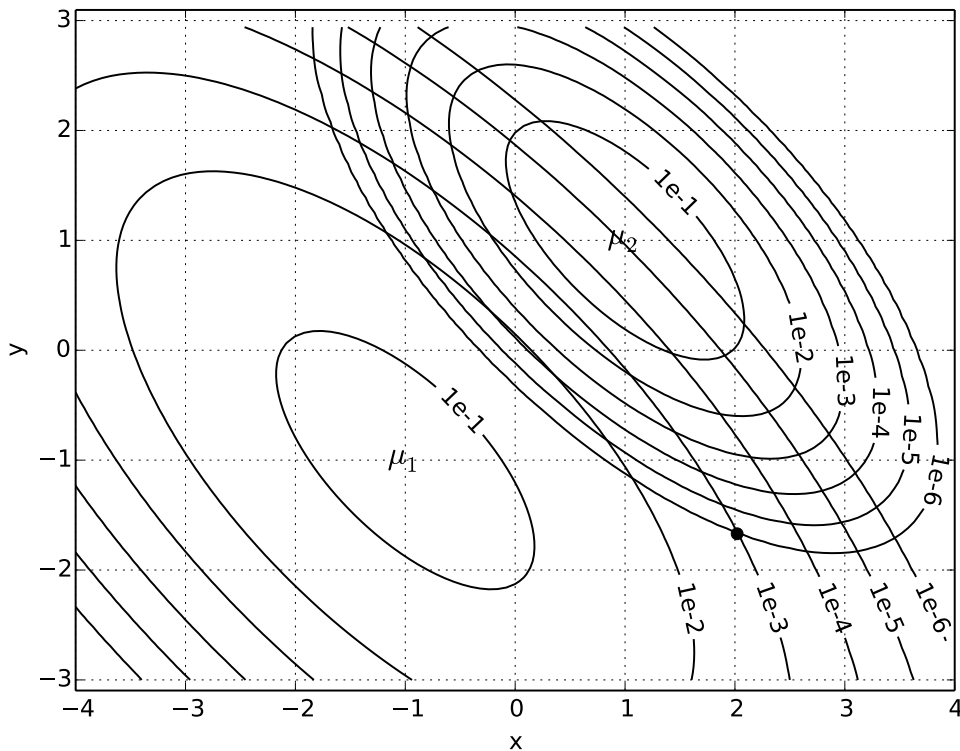


Figure 1. The two probability distribution functions (PDFs) from problem B-3. The mean values are indicated by μ_1 and μ_2 . The contour lines correspond to the locations where the PDFs equal the following values: $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$.

B-3 Probability based learning

(4p)

Figure 1 displays a two-class Maximum Likelihood (ML) classifier where the likelihood of a data point x given each class c is a two dimensional multivariate Gaussian (Normal) distribution,

$$P(x|c) = \mathcal{N}(x, \mu_c, \Sigma_c).$$

The mean vectors for class 1 and 2 are respectively

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- a) looking at the figure, determine which of the two covariance matrices below corresponds to class 1 and 2 respectively, give a qualitative explanation for your answer:

$$\Sigma_a = \begin{bmatrix} 0.3 & -0.2 \\ -0.2 & 0.3 \end{bmatrix} \quad \Sigma_b = \begin{bmatrix} 1.2 & -0.8 \\ -0.8 & 1.2 \end{bmatrix},$$

- b) using Figure 1, classify the following points according to the ML classifier:

$$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad x_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad x_4 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

- c) is the separation boundary between the two classes linear (straight line)? Motivate your answer using Figure 1.
- d) You want to classify the point indicated by the black circle in Figure 1 with a Maximum a Posteriori (MAP) classifier based on the likelihood functions above and the a priori probabilities of the classes $p_1 = P(c = 1)$ and $p_2 = P(c = 2)$. What is the range for p_1 and p_2 that will make the MAP classifier assign the point to class 2?

Note: no complex derivations are required to solve any of the above questions. If you need to refer to the figure to explain your solution, **do not draw on this sheet:** we will not receive it with your solutions.

B-4 Information Content

(3p)

Imagine that you are playing with Cards and randomly sample three cards out of the pile of 52 cards *with replacement* (i.e. you sequentially draw a card but return it to the pile each time you have seen what it is).

- a) At each instance of drawing a card, what is the Shannon information content of the outcome with respect to the suit, one of {*Clubs*, *Spades*, *Diamonds*, *Hearts*}, measured in bits?
- b) You play a game with a rule that you win if all the *three* cards are of *different* suit from each other. Otherwise you lose. How unpredictable is the outcome of this game (win or lose)? Answer in terms of entropy, measured in bits.
- c) With respect to the outcome of the game in **b)**, what is the expected information gain from seeing the suit of the first card?
- d) Now, what is the expected information gain from also seeing the suit of the second card?

Note: if you do not have a calculator, answer with an expression but simplify it as much as possible.

B-5 Ensemble Methods

(2p)

Briefly answer the following questions regarding ensemble methods of classification.

- a) What are the two kinds of randomness involved in the design of Decision Forests?
- b) In Adaboost algorithm, each training sample is given a weight and it is updated according to some factors through an iteration of training weak classifiers. What are the two most dominant factors in updating the weights, and how are they taken into account?

B-6 Regression with regularization

(3p)

For a set of N training samples $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, each consisting of input vector \mathbf{x} and output y , suppose we estimate the regression coefficients $\mathbf{w}^\top (\in \mathbf{R}^d) = \{w_1, \dots, w_d\}$ in a linear regression model by minimizing

$$\sum_{n=1}^N (y_n - \mathbf{w}^\top \mathbf{x}_n)^2 + \lambda \sum_{i=1}^d w_i^2$$

for a particular value of λ .

For parts **a)** through **d)**, indicate which of i. through v. is correct. Briefly justify your answer.

- a)** As we increase λ from 0, the training error (residual sum of squares, RSS) will:
- i. Steadily increase.
 - ii. Steadily decrease.
 - iii. Remain constant.
 - iv. Increase initially, and then eventually start decreasing in an inverted U shape.
 - v. Decrease initially, and then eventually start increasing in a U shape.
- b)** Repeat **a)** for test RSS.
- c)** Repeat **a)** for variance.
- d)** Repeat **a)** for (squared) bias.

B-7 Support Vector Machines

(3p)

Training a support vector machine using a quadratic kernel

$$\mathcal{K}(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + 1)^2$$

has resulted in the following three support vectors:

$$s_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad s_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \quad s_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

The first is a positive sample with $\alpha = 0.25$ while the other two are negative samples where $\alpha = 0.1$.

- a)** Determine how the following *new* datapoints will be classified:

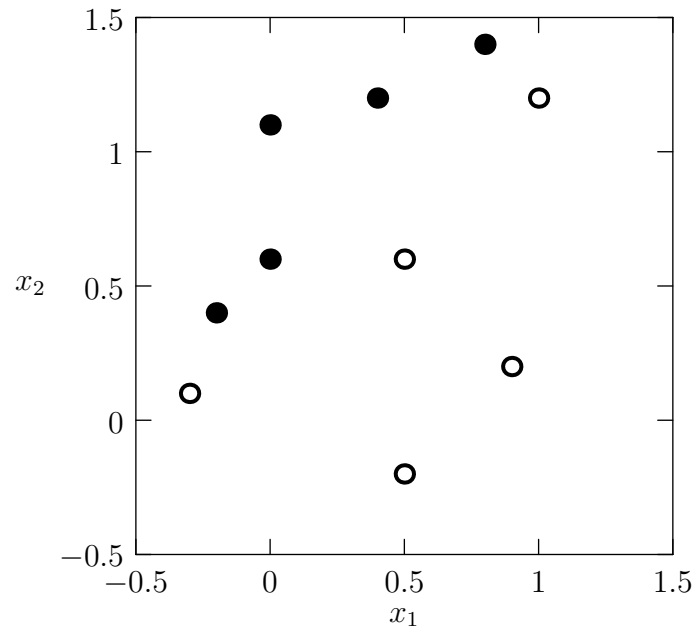
$$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad x_2 = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad x_4 = \begin{bmatrix} 2.5 \\ 1 \end{bmatrix}$$

You must show the formulas and calculations used to arrive at your answer.

- b)** Draw a diagram of the input space and show the shape and position of the decision boundary.

B-8 Linear classifier

(3p)



- a) The figure illustrates a dataset where filled circles are positive samples and unfilled are negative. For a single “artificial neuron”, find a set of values for the weights and the threshold that will separate the positive from the negative samples.

Note: You do not have to use any learning algorithm, but you must explain where you got the values from.

- b) If a new negative sample at $(0.5, 0.0)$ arrives, and learning is done with the *perceptron learning rule*; how will these values change?