KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2020-03-13, kl 14.00 – 18.00

Aids allowed: *calculator*, *language dictionary*.

In order to pass, you have to fulfill the requirements defined both in the A- and in the B-section.

A Questions on essential concepts

Note: As a prerequisite for passing you must give the correct answer to almost *all* questions. Only *one* error will be accepted, so pay good attention here.

A-1 Shannon Entropy

Consider a single toss of fair coin. Regarding the uncertainty of the outcome {head, tail}:

a) The entropy is equal to two bits.

b) The entropy is equal to one bit.

c) The entropy is not related to uncertainty.

Solution: b

A-2 Regression and Classification

Choose the most proper statement reflecting the output formats of regression and classification.

a) They are both discrete.

b) Discrete for classification and continuous-valued for regression.

c) Discrete for regression and continuous-valued for classification.

Solution: b

A-3 Probabilistic Learning

Which of the following statements is *false*?

a) *Probabilistic learning* involves estimating $P(\mathbf{x}, y)$ from a dataset $\mathcal{D} = \{(\mathbf{x}, y)_1, \ldots, (\mathbf{x}, y)_n\}$.

b) *Probabilistic learning* helps one work with uncertainty in a problem domain.

c) *Probabilistic learning* can only be used to create generative models.

Solution: c

A-4 Maximum Likelihood Estimation

Given a dataset $\mathcal{D} = \{(\mathbf{x}, y)_1, \ldots, (\mathbf{x}, y)_n\}$, Maximum Likelihood estimates of the parameters of $P(\mathbf{x}, y)$ are computed with which assumption and optimality criterion?

a) Choose the parameters that maximize the likelihood of $\mathcal{D}$ assuming all observations of $\mathcal{D}$ are independently and identically distributed multivariate Gaussian.

b) Choose the parameters that maximize the likelihood of $\mathcal{D}$ assuming all observations of $\mathcal{D}$ are independently and identically distributed given y.

c) Choose the parameters that maximize the likelihood of $\mathcal{D}$ assuming it is a representative sample of the problem domain.

Solution: b

A-5 Artificial Neural Networks

Which statement describes the functionality of an artificial neuron (the *perceptron*)?

a) The perceptron generates an output signal based on the integrated weighted input.

b) Each perceptron solves a partial differential equation.

c) The perceptron can be trained to compute arbitrary complex functions.

Solution: a

A-6 Support Vector Machine

What does the concept of *Structural Risk Minimization* address?

a) Splitting the data set such that training and testing is supported.

b) Selecting a separating hyperplane such that future data is most likely classified correctly.

c) Exploring multiple training methods to identify the best classification.

Solution: b

A-7 Ensemble Learning

Which one below *correctly* describes the characteristics of boosting method in machine learning?

a) Each training example has a weight which is re-weighted through iterations.

b) Weak classifiers to be combined are chosen independently of each other.

c) Weak classifiers are ensembled with equal contributions (reliablilty).

Solution: a

A-8 Principal Component Analysis (PCA)

To appply the maximum variance criterion in the Principal Component Analysis (PCA),

a) covariance matrix of given vectors

b) random sampling from given vectors

c) least squares approximation

is a useful tool. Choose the most proper statement.

Solution: a

Note: Your answers (eight of them) need be on one solution sheet (**we will not receive this page!**).

B Graded problems

A pass is guaranteed with the required points for 'E' below in this section *and* the prerequisite in the A-section.

Preliminary number of points required for different grades:

$$
\begin{aligned}
24 \leq p \leq 27 &\rightarrow A \\
20 \leq p < 24 &\rightarrow B \\
16 \leq p < 20 &\rightarrow C \\
12 \leq p < 16 &\rightarrow D \\
9 \leq p < 12 &\rightarrow E \\
0 \leq p < 9 &\rightarrow F
\end{aligned}
$$

B-1 Terminology (4p)

For each term (a–h) in the left list find the explanation that *best* describes how the term is used in machine learning among the list in the right, and indicate it by the number.

a) Expectation Maximization

b) Posterior probability

c) RANSAC

d) Dropout

e) The Lasso

f) Bagging

g) Error back-propagation

h) k-fold cross validation

1) A technique for assessing a model while exploiting available data for training and testing

2) A method for preventing artificial neural networks from overfitting

3) Algorithm to learn with latent variables

4) The last solution

5) Conditional probability taking into account the evidence

6) Probability at a later time

7) An approach to regression that results in feature seletion

8) Sudden drop of performance

9) A strategey to generate k different models

10) Random strategy for amplitude compensation

11) Algorithm to train artificial neural networks

12) Implementation of the bag-of-words model

13) Estimating expected value

14) Bootstrap aggregating

15) Robust method to fit a model to data with outliers

Solution: a-3, b-5, c-15, d-2, e-7, f-14, g-11, h-1

B-2 Decision Trees and Decision Forests (3p)

a) Draw a decision tree of depth two that is consistent with the training data in the table. $a_1, ..., a_4$ are attributes of the training data. (It is not required to use information gain to solve the task.) (1p)

a_1	a_2	a_3	a_4	$class$
0	0	0	0	-
0	0	1	1	+
0	1	0	1	+
0	1	1	0	-
1	0	0	0	+
1	0	1	1	+
1	1	1	0	-
1	1	0	1	-

b) Mainly two kinds of randomness are known to form the principle of Decision Forests.

In which two of the following processes are those randomnesses involved? (1p)

i. In deciding the number of trees used.
ii. In deciding the depth of trees used.
iii. In the way to formulate the information gain.
iv. In feature selection at each node.
v. In generating bootstrap replicas.

Simply indicate your choices.

c) Suppose we have generated a Decision Forest using five bootstrapped samples from a data set containing three classes, {Blue, Yellow, Red}. We then applied the forest to a specific test input, x, and observed five estimates of $P(\text{Class is Yellow}|x)$: 0.3, 0.35, 0.55, 0.7, and 0.75.

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach, and the other is based on the average probability. In this example, what is the final classification under each of these two approaches? (1p)

i. Blue in majority vote and Yellow in averaging.
ii. Blue or Red in majority vote and Yellow in averaging.
iii. Blue or Red in both approaches.
iv. Yellow in both approaches.
v. Yellow in majority vote and Red in averaging.

Motivate your answer by short phrases.

Solution:

b)-iv and v

c)-iv. The *Yellow* class has three votes in majority vote, and the average probability, 0.53, is higher than those of other classes.

B-3 Nearest Neighbor, Classification (3p)

Suppose that we take a data set, divide it into training and test sets, and then try out two different classification procedures. We use half of the data for training, and the remaining half for testing.

First we use k-nearest neighbor (where $k = 1$) and get an average error rate (averaged over both test and training data sets) of 6%. Next we use Logistic Regression and get an error rate of 8% on the training data. We also get the average error rate (averaged over both test and training data sets) of 9%.

a) What was the error rate with 1-nearest neighbor on the test set? (1p)

b) What was the error rate with the Logistic Regression on the test set? (1p)

c) Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning. (1p)

Solution:

b) 12%. Training error for 1-NN is always zero, and therefore the testing error is 12%.

a) 10%.

c) Logistic Regression because it achieves lower error rate on the test data (10% < 12%).

B-4 Regression with regularization (3p)

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

Now, let us consider different models trained with different values of λ, starting from a very large value (infinity) and *decreasing* it down to 0. Then, for parts **a**) through **c**), indicate which of i. through v. is correct. (1p, each)

Briefly justify each of your answers.

a) As we decrease λ, the variance of the model 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 the training error (resisual sum of squares, RSS).

c) Repeat **a**) for test RSS.

Solution: **a**)-i, **b**)-ii, **c**)-v
If λ was infitity, all w_i would be zero; the model is extremely simple, predicting a constant and has no variance with the prediction being far from actual value (thus with high bias). As we decrease λ, all w_i increase from zero toward their least square estimate values while steadily decreasing the *training* error to the ordinary least square RSS (and also decreasing bias) as the model continues to better fit training data. The values of w_i then become more dependent on training data, thus increasing the variance. The *test* error initially decreases as well, but eventually start increasing due to overfitting to the training data.

B-5 The Subspace Method (2p)

Given a set of feature vectors which all belong to a specific class C (i.e. with an identical class label), we performed PCA on them and generated an orthonormal basis $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ as the outcome. When the training samples in the class are well localised, the basis can be considered as a tool to represent possible variations of feature vectors within C in terms of a p-dimensional subsapce, $\mathcal{L}$.

Provide an answer to the following questions.

a) We have a new input vector $\mathbf{x}$ whose class is unknown, and consider its projectiton length on $\mathcal{L}$. Describe how the projection length is represented, using a simple formula. (1p)

b) Now, we consider to solve a K-class classification problem with the Subspace Method and assume that a subspace $\mathcal{L}^{(j)}$ $(j = 1, ..., K)$ has been computed with training data for each class, respectively. Briefly explain the way to determine the class to which vector $\mathbf{x}$ should belong using the projection lengths. (1p)

Solution:

a) $\sqrt{S}$ where $S = \Sigma_{i=1}^{p}(\mathbf{x}, \mathbf{u}_i)^2$

b) $\mathbf{x}$ should belong to the class where the projection length to the corresponding subspace is maximised.

B-6 Probability based learning (3p)

Using the dataset $\mathcal{D}$ shown in Table 1, will I go running on a cool day that has a rainy outlook, high humidity, and no wind?

a) Answer this using Maximum Likelihood classification. (1p)

b) Answer this using Maximum a Posteriori classification. (1p)

c) Answer this using Naive Bayes classification. (1p)

n	$\mathbf{x}_n$				y_n
day	outlook	temperature	humidity	windy	run
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	high	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	cool	high	false	no

Table 1. Dataset $\mathcal{D}$ showing weather conditions when I did or did not go running.

Solution:

a) For maximum likelihood classification, we find the class according to:

$$y_{\text{ML}} = \arg \max_{y \in \{\text{yes,no}\}} P(\mathbf{x}|Y = y, \mathcal{D})$$

For $\mathbf{x} = (\text{rainy}, \text{cool}, \text{high}, \text{false})$, we need to compare $P(\mathbf{x}|Y = \text{yes}, \mathcal{D})$ and $P(\mathbf{x}|Y = \text{no}, \mathcal{D})$. Of the nine instances where $Y = \text{yes}$ in the dataset, only one shows this $\mathbf{x}$ and so:

$$P(\mathbf{x}|Y = \text{yes}, \mathcal{D}) = 1/9$$
$$P(\mathbf{x}|Y = \text{no}, \mathcal{D}) = 1/5.$$

Maximum likelihood classification then says $Y = \text{no}$.

b) For maximum a posteriori classification, we find the class according to:

$$y_{\text{MAP}} = \arg \max_{y \in \{\text{yes,no}\}} P(\mathbf{x}|Y = y, \mathcal{D}) P(Y = y|\mathcal{D})$$

We can use the results from the previous part, and compute the priors:

$$P(Y = \text{yes}|\mathcal{D}) = 9/14$$
$$P(Y = \text{no}|\mathcal{D}) = 5/14$$

Thus

$$P(\mathbf{x}|Y = \text{yes}, \mathcal{D}) P(Y = \text{yes}|\mathcal{D}) = 1/9 \cdot 9/14 = 1/14$$
$$P(\mathbf{x}|Y = \text{no}, \mathcal{D}) P(Y = \text{no}|\mathcal{D}) = 1/5 \cdot 5/14 = 1/14.$$

So MAP classification says it's a toss-up.

c) Using Naive Bayes, we find the class according to:

$$y_{\text{NB}} = \arg\max_{y \in \{\text{yes,no}\}} P(Y = y|\mathcal{D}) \prod_{i=1}^{4} P([\mathbf{x}]_i|Y = y, \mathcal{D})$$

Now we just have to compute:

$$\begin{aligned}
P([\mathbf{x}]_1 = \text{rainy}|Y = \text{yes}, \mathcal{D}) &= 3/9 \\
P([\mathbf{x}]_2 = \text{cool}|Y = \text{yes}, \mathcal{D}) &= 3/9 \\
P([\mathbf{x}]_3 = \text{high}|Y = \text{yes}, \mathcal{D}) &= 4/9 \\
P([\mathbf{x}]_4 = \text{false}|Y = \text{yes}, \mathcal{D}) &= 6/9 \\
P([\mathbf{x}]_1 = \text{rainy}|Y = \text{no}, \mathcal{D}) &= 2/5 \\
P([\mathbf{x}]_2 = \text{cool}|Y = \text{no}, \mathcal{D}) &= 2/5 \\
P([\mathbf{x}]_3 = \text{high}|Y = \text{no}, \mathcal{D}) &= 4/5 \\
P([\mathbf{x}]_4 = \text{false}|Y = \text{no}, \mathcal{D}) &= 3/5
\end{aligned}$$

Putting this all together:

$$P(Y = \text{yes}|\mathcal{D}) \prod_{i=1}^{4} P([\mathbf{x}]_i|Y = \text{yes}, \mathcal{D}) = (9/14)(3/9)(3/9)(4/9)(6/9) = 0.021$$

$$P(Y = \text{no}|\mathcal{D}) \prod_{i=1}^{4} P([\mathbf{x}]_i|Y = \text{no}, \mathcal{D}) = (5/14)(2/5)(2/5)(4/5)(3/5) = 0.027$$

So according to Naive Bayes, the answer is 'no'.

B-7 Probability based Learning (3p)

You are told that the elements of the dataset $\mathcal{D} = \{-1, 1, 0\}$ were independently drawn from a univariate Gaussian distribution with mean $\mu = -1$ and variance $\sigma^2 = 2$.

a) Compute the likelihood of $\mathcal{D}$ in this model. (0.5p)

b) Estimate the parameters of a univariate Gaussian for $\mathcal{D}$ using the maximum likelihood optimality criterion. (1p)

c) Consider that a fourth point is observed. Find all values of this point such that the likelihood of the four observations is greater in the univariate Gaussian distribution with mean $\mu = -1$ and variance $\sigma^2 = 2$ than in the distribution your found in part b. (1.5pt)

Solution:

a) With these parameters, the univariate Gaussian distribution is

$$Pr(x|\theta) = \frac{1}{\sqrt{\pi 4}} e^{-(x+1)^2/4}.$$

Then the likelihood of $\mathcal{D}$ in this model is

$$\begin{aligned}
Pr(\mathcal{D}|\theta) &= \prod_{n=1}^{3} \frac{1}{\sqrt{\pi 4}} e^{-(x_n+1)^2/4} = \left(\frac{1}{\sqrt{\pi 4}}\right)^3 e^{-(0)^2/4} e^{-(1)^2/4} e^{-(2)^2/4} \\
&= \left(\frac{1}{\sqrt{\pi 4}}\right)^3 e^{-5/4} < 0.0064.
\end{aligned}$$

b) The maximum likelihood estimation of the parameters is computed using:

$$\begin{aligned}\theta_{\text{ML}} &= \arg\max_{\theta} P(\mathcal{D}|\theta) = \arg\max_{\theta} \log P(\mathcal{D}|\theta) \\ &= \arg\max_{\theta=\{\mu,\sigma^2\}} \sum_{n=1}^{3} \log\left(\frac{1}{\sqrt{\pi 2\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}\right) \\ &= \arg\max_{\theta=\{\mu,\sigma^2\}} -0.5(3)\log\sigma^2 - \sum_{n=1}^{3}\frac{(x_n-\mu)^2}{2\sigma^2}\end{aligned}$$

Taking the partial derivative of this expression with respect to μ produces the ML estimate of that parameter:

$$\mu_{\text{ML}} = \frac{1}{3}\sum_{n=1}^{3} x_n = \frac{1}{3}(1-1+0) = 0.$$

Taking the partial derivative of this expression with respect to σ^2 setting $\mu = 0$ produces:

$$\sigma^2_{\text{ML}} = \frac{1}{3}\sum_{n=1}^{3}(x_n)^2 = \frac{2}{3}.$$

c) The likelihood of $\mathcal{D}$ in the ML model is

$$\begin{aligned}Pr(\mathcal{D}|\theta_{\text{ML}}) &= \prod_{n=1}^{3}\frac{1}{\sqrt{\pi 2(2/3)}} e^{-(x_n)^2/(4/3)} \\ &= \left(\frac{1}{\sqrt{\pi 4/3}}\right)^3 e^{-(0)^2/(4/3)} e^{-(1)^2/(4/3)} e^{-(-1)^2/(4/3)} \\ &= \left(\frac{1}{\sqrt{\pi 4/3}}\right)^3 e^{-3/2} < 0.0261.\end{aligned}$$

Using the result of part a, we need to find the range of x that satisfy:

$$\left(\frac{1}{\sqrt{\pi 4}}\right)^4 e^{-5/4} e^{-(x+1)^2/4} > \left(\frac{1}{\sqrt{\pi 4/3}}\right)^4 e^{-3/2} e^{-x^2/(4/3)}$$

Simplifying this

$$\begin{aligned}\left(\frac{1}{\sqrt{\pi 4}}\right)^4 e^{-5/4} e^{-(x+1)^2/4} &> \left(\frac{1}{\sqrt{\pi 4/3}}\right)^4 e^{-3/2} e^{-x^2/(4/3)} \\ e^{x^2/(4/3)} e^{-(x+1)^2/4} &> \left(\frac{\sqrt{\pi 4}}{\sqrt{\pi 4/3}}\right)^4 e^{-1/4} \\ x^2/(4/3) - (x+1)^2/4 &> -1/4 + 4\log(4/(4/3)) \\ 3x^2 - (x+1)^2 &> -1 + 16\log(3).\end{aligned}$$

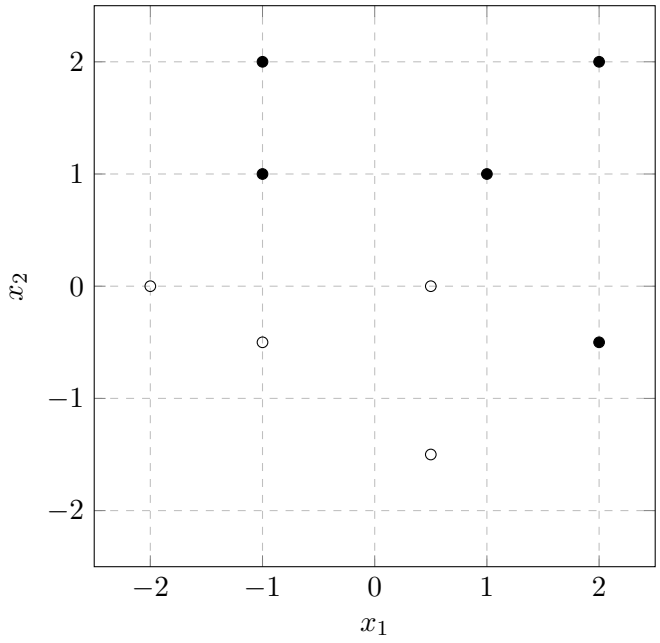
So the quadratic equation to solve is:

$$2x^2 - 2x - 16\log(3) = 0.$$

The roots of this quadratic equation are $x = 2.52$ and $x = -1.52$. If $x \in [-1.52, 2.52]$, then the likelihood of the dataset in the maximum likelihood model is greater. So x must be outside this range for the likelihood in the other model to be higher.

B-8 Support Vector Machines (3p)

Given the training data illustrated in the figure. Filled circles are positive examples, unfilled are negative.



a) The data is linearly separable, so a linear kernel should be sufficient. Why could it be reasonable to use a quadratic kernel anyway? (1p)

b) If a linear kernel is used, what will the support vectors be? (1p)

c) If a quadratic kernel is used, state at least one of the data points which will surely not be a support vector. Motivate with a short argument why this point is unlikely to be a support vector. (1p)

Solution:

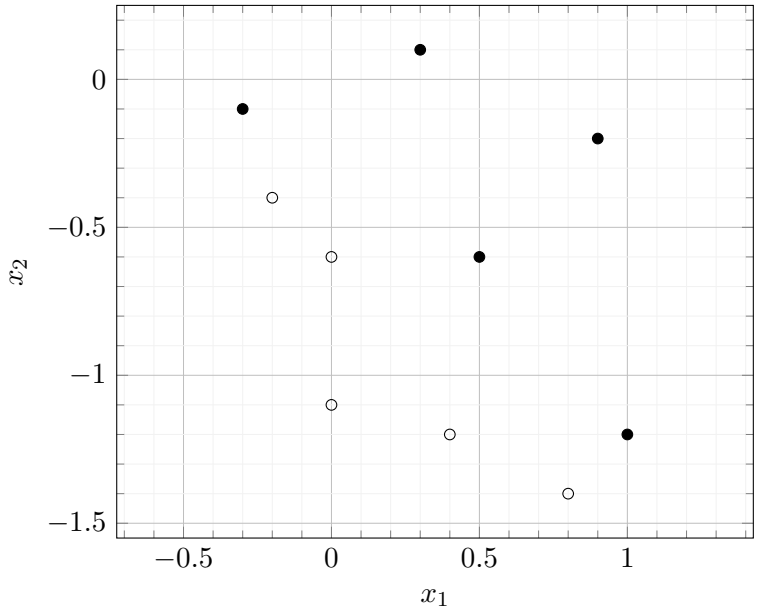
a) A quadratic kernel will result in much wider margins

b) $(-1, 1), (2, -0.5), (0.5, 0)$

c) $(-1, 2), (-1, -0.5), (2, 2)$ will not be support vectors.

All of these have another point of the same class much closer to the boundary between the classes.

B-9 Perceptron Learning, Linear Classification (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. Make sure that positive points are *above* the threshold and negative points are *below*. (2p)

Note: There is no need to use any learning algorithm here. It is sufficient to find a suitable linear separator in the figure, but you must explicitly state the weight and threshold values and explain *in keywords* how you arrived at these values.

b) If a new positive sample at $(0.5, -0.5)$ arrives, and learning is done with the *perceptron learning rule*; how will these values change? (1p)

Solution:

a) A natural separator is the line passing through $(-0.5, 0.0)$ and $(1.0, -1.5)$. The weight vector is always perpendicular to this line and points towards the positive sample side (here: upper right).

$$w_1 = 1, \qquad w_2 = 1$$

We get the threshold, θ, by testing a point on the line, for example $(0.0, -0.5)$, using the weights we have chosen.

$$0.0 \cdot w_1 + 0.5 \cdot w_2 - \theta = 0 \quad \Rightarrow \quad \theta = 0.5$$

b) No change, because the sample is already correctly classified.