KTH Computer Science
and Communication

Exam in DD2421 Machine Learning
2019-10-25, kl 08.00 – 12.00

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.

Note: Your answers from section A (eight of them) must be written on a *single* solution sheet. **We will not receive this question page**.

A-1 Regression and Classification

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

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

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

c) They are both continuous-valued.

Solution: a

A-2 Shannon Entropy

Consider a single toss of *skewed* coin (it is likely to show one side more than the other side). Regarding the uncertainty of the outcome {head, tail}:

a) The entropy is smaller than one bit.

b) The entropy is equal to two bits.

c) The entropy does not explain the uncertainty.

Solution: a

A-3 Probabilistic Learning

What is the goal of *maximum likelihood* classification of an observation?

To find the class that:

a) maximizes the probability of the observation conditioned on the class.

b) has the maximum probability in a Gaussian distribution.

c) has the maximum prior probability.

Solution: a

A-4 Naive Bayes Classifier

Naive Bayes classification assumes $Pr(x_1, \ldots, x_D \,|\, Y = y) = \prod_{d=1}^{D} Pr(x_d \,|\, Y = y)$. This assumption means:

a) All D dimensions of an observation are conditionally distributed Bernoulli.

b) All D dimensions of an observation are conditionally independent given Y.

c) Y is conditionally independent of all D dimensions of an observation.

Solution: b

A-5 Perceptron Learning

When does the *perceptron learning algorithm* stop modifying the weights?

a) When the step size reaches zero.

b) When the error gradient becomes zero.

c) When all training data is correctly classified.

Solution: c

A-6 Support Vector Machine

When using a *kernel*-function, for example in a support vector machine; what does this function correspond to, mathematically?

a) The generalised distance between any data point and the decision boundary.

b) The scalar product between two data points transformed into a higher dimensional space.

c) The midpoint of the training data, computed separately for each class.

Solution: b

A-7 Ensemble Learning

Which one below correctly describes the property of *Adaboost Algorithm* for classification?

a) Models to be combined are requied to be as similar as possible to each other.

b) A weight is given to each training sample, and it is iteratively updated.

c) Adaboost algorithm is more suited to multi-class classification than binary classification.

Solution: b

A-8 Principal Component Analysis (PCA)

Which one is considered as the main purpose of the principal component analysis (PCA)?

a) To find the least squares fit.

b) To treat the data in infinite dimensional space.

c) To reduce the effective number of variables.

Solution: c

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:

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

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.

a) k-means

b) The Lasso

c) RANSAC

d) Dropout

e) Curse of dimensionality

f) k-nearest neighbour

g) Expectation Maximization

h) k-fold cross validation

1) An approach to train artificial neural networks

2) Estimating expected value

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

4) Random strategy for amplitude compensation

5) An approach to regression that results in feature seletion

6) The final solution

7) Method for estimating the mean of k observations

8) Clustering method based on centroids

9) Sudden drop of performance

10) Issues in data sparsity in space

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

12) An approach to generate k different models

13) Algorithm to learn with latent variables

14) Problems in slow computation

15) Class prediction by a majority vote

Solution: a-8, b-5, c-3, d-1, e-10, f-15, g-13, h-11

B-2 Classification (3p)

Suppose that we take a data set, divide it into two parts of equal size, Part I and Part II. We try out two different classification procedures, by using Part I and Part II as our training set and test set, respectively, That is, we use half of the data for training, and the remaining half for testing.

a) First we use 1-Nearest Neighbor rule (1-NN) and get an average error rate (averaged over both test and training data sets) of 7%. What was the error rate with 1-nearest neighbor on the test set? Briefly reason the answer.

b) Next we use the Adaboost Algorithm and get an error rate of 10% on the training data. We also get the average error rate (averaged over both test and training data sets) of 11%. What was the error rate with the Adaboost Algorithm on the test set?

c) Now, we swap the roles of Part I and Part II, and repeat the same experiments. On the test set (Part I), we get an error rate of 10% with both 1-NN and the Adaboost Algorithm. Based on all these results, by the cross-validation, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

Solution:

a) 14%. (Training error for 1-NN is always zero, and therefore the testing error is 14%.)

b) 12%.

c) Adaboost because it achieves lower error rate on the test data on average (11% < 12%).

B-3 Decision Forests (Random Forests) (3p)

a) Training a decision tree, a component of Decision Forests, involves generating questions to be asked at different nodes such that the expected information gain is maxmised in terms of entropy. Formulate the entropy E in the class distribution at a node, using p_i as the probability for class i. (Simply consider the entropy before splitting the data at the node.)

b) Mainly two kinds of randomness are known to form the basic principle of Decision Forests. In which two of the following processes are those randomnesses involved?

i. In the rule of terminating a node as a leaf node.
ii. In the way to formulate the information gain.
iii. In feature selection at each node.
iv. In deciding the number of trees used.
v. In generating bootstrap replicas.
vi. In combining the results from multiple trees.

Simply indicate two among those above.

c) Suppose we have generated a Decision Forest using five bootstrapped samples from a data set containing three classes, {Red, Yellow, Green}. We applied it to a specific test input, $\mathbf{x}$, and observed five estimates of $P(\text{Class is Green}|\mathbf{x})$: 0.80, 0.65, 0.45, 0.70, and 0.40.

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

i. Green in both approaches.
ii. Yellow or Red in both approaches.
iii. Yellow or Red in averaging and Green in majority vote.

Indicate one among the above, and motivate your answer by short phrases.

Solution:

a) Entropy $= \sum_i -p_i \log_2 p_i$

b) iii and v.

c) i.
The *Green* class has three votes in majority vote, and the average probability, 0.60, must be higher than any of the other classes.

B-4 Bias and Variance (3p)

a) One of the four subfigures (i)-(iv) in Figure 1 displays the typical trend of prediction error of a model for training and testing data with comments on its bias and variance, {high, low}. Simply indicate one of the four figures which most well represents the general situation.

What is the main reason for the prediction errors on the red curve to appear in a U-shape?

b) Now consider the specific case of using *Bagging* by an ensemble of decision tree classifiers. What sort of improvememt can be expected in the ensemble predictions in terms of *bias* or *variance* of the classifier as a whole?

c) In ridge regression, relative to least squares, a term called *shrinkage penalty* is added in the quantity to be minimised. They give improved prediction accuracy in some situations. Briefly explain when this happens in terms of bias-variance trade-off.

Solution:

a) (iii)
Overfitting.

b) Reduction of the variance.

c) When the increase in bias is less than the decrease in variance.

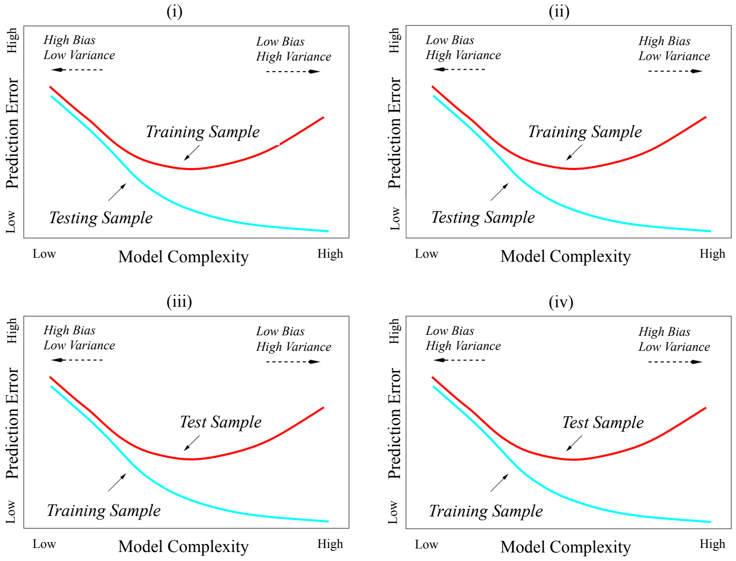


Figure 1. Graphs for Problem B-4.

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.

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.

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)

Consider you have a dataset of N independent and identically distributed labeled observations, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. You choose the following conditional probability model for regression:

$$Y|X = x \sim \mathcal{N}(\log(w|x|), 1) \qquad (1)$$

that is, the dependent variable is conditionally distributed Gaussian with a mean $\log(w|x|)$ and variance 1:

$$Pr(y|x,w) = \frac{1}{\sqrt{2\pi}} e^{-(y-\log(w|x|))^2/2} \tag{2}$$

Find the maximum likelihood estimate of the parameter w given the dataset.

Solution: The maximum likelihood estimate of w maximizes the likelihood of the dataset, i.e.,

$$w_{\text{ML}} := \arg\max_w \prod_{n=1}^{N} Pr(y_n|x_n,w) = \arg\max_w \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(y_n-\log w|x_n|)^2/2} \tag{3}$$

Since the log is monotonically increasing, the above can be written

$$w_{\text{ML}} := \arg\max_w \sum_{n=1}^{N} -(y_n - \log w|x_n|)^2 = \arg\min_w \sum_{n=1}^{N} (y_n - \log w|x_n|)^2. \tag{4}$$

Taking the derivative of the expression with respect to w and setting the result to zero produces

$$\sum_{n=1}^{N} y_n = \sum_{n=1}^{N} \log w_{\text{ML}}|x_n| = N \log w_{\text{ML}} + \sum_{n=1}^{N} \log|x_n|. \tag{5}$$

Hence the ML estimate of the parameter is

$$w_{\text{ML}} = \exp\left[\frac{1}{N}\sum_{n=1}^{N}(y_n - \log|x_n|)\right]. \tag{6}$$

B-7 Probability based Learning (3p)

Consider the data in Table 1, showing the conditions of each day the past two weeks that I did or did not run around Hagaparken. Create a Naive Bayes classifier to determine if I will go run given that the weather outlook today (day 15) is rainy, the temperature is cool, the humidity is high, and it is not windy. In other words, find the y that maximizes $P(y|\mathbf{x}_{15})$. Motivate all your decisions. (Hint: Use Bayes to rewrite $P(y|\mathbf{x}_{15})$, then apply the Naive Bayes assumption, then estimate all values from the given dataset.)

Solution: We want to find the y that maximizes $P(y|\mathbf{x}_{15})$. This is also the y that maximizes $P(\mathbf{x}_{15}|y)P(y)$ by Bayes' rule. We can easily estimate the priors from the dataset: $Pr(\text{yes}) = 9/14$, and $Pr(\text{no}) = 5/14$. Naive Bayes classification assumes all dimensions are conditionally independent given the dependent variable. Thus

$$P(\mathbf{x}|y) = P(x_1|y)P(x_2|y)P(x_3|y)P(x_4|y). \tag{7}$$

We just have to estimate for each possible y:

$$P(\mathbf{x}|y) = P(\text{rainy}|y)P(\text{cool}|y)P(\text{high}|y)P(\text{false}|y). \tag{8}$$

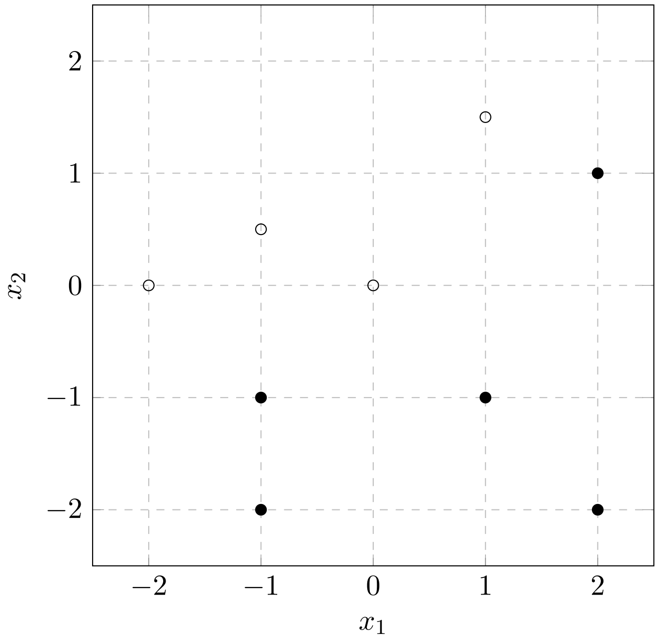
From the table, $P(\text{rainy}|\text{yes}) = 3/9$; $P(\text{cool}|\text{yes}) = 3/9$; $P(\text{high}|\text{yes}) = 3/9$; and $P(\text{false}|\text{yes}) = 6/9$. Multiplying the prior $Pr(\text{yes})$ by these gives $1458/91854 = 1/63 = 0.016$. Now for not going running, the table shows: $P(\text{rainy}|\text{no}) = 2/5$; $P(\text{cool}|\text{no}) = 1/5$; $P(\text{high}|\text{no}) = 4/5$; and $P(\text{false}|\text{no}) = 2/5$. Multiplying the prior $Pr(\text{no})$ by these gives $80/8750 = 8/875 = 0.009$. Since the larger value is for going running, Naive Bayes classifies the day as going running.

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	normal	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	mild	high	true	no
15	rainy	cool	high	false	?

Table 1. Dataset showing weather conditions when I did or did not go running.

B-8 Support Vector Machines (4p)

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



a) If a linear kernel is used, what will the support vectors be?

b) Which of the support vectors in **a** would have the largest associated α-value? You need to motivate the answer!

c) The data is linearly separable, so a linear kernel should be sufficient. What would be the advantage of using a non-linear kernel anyway?

d) If a quadratic kernel and no slack 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.

Solution:

a) The support vectors will be $[-1, -1]$, $[0, 0]$, and $[2, 1]$.

b) $[0, 0]$ will have the largest α. This is the only negative support vector, so its α-value has to balance the sum of the two positive support vectors.

c) A non-linear kernel will result in much *wider margins* and therefore *generalize better*.

d) $[-1, -2]$, $[2, -2]$, and $[-1, 0.5]$ will not be support vectors.

All of these have another point from the same class much closer to the boundary between the classes, preventing the margin to extend out to these points.

B-9 Artificial Neural Networks (2p)

Consider the training data in the table for a classification task where $+$ means a positive sample and $-$ a negative. Each row represents one training sample.

x_1	x_2	Class
0	1	$+$
1	0	$+$
2	2	$-$
2	-1	$-$
-1	-1	$-$
-1	2	$-$

a) What is the *minimum* number of *layers* needed for a feed-forward artificial neural network to correctly classify all these points? You must motivate your answer.

b) What should the number of input and output nodes be in this network?

Solution:

a) Two layers is needed and sufficient. The points are not linearly separable so one layer is not sufficient. Two layers can solve any separation problem.

b) Two input nodes (corresponding to x_1 and x_2) and one output node (corresponding to the class).