KTH Computer Science and Communication

Exam in DD2421 Machine Learning
2023-10-26, kl 08.00 – 12.00

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

To take this exam you must be registered to this specific exam as well as to the course.

In order to pass this exam, your score x first needs to be 20 or more (out of 42, full point). In addition, given your points y from the Programming Challenge (out of 18, full point), the requirements on the total points, $p = x + y$, are preliminarily set for different grades as:

$$
\begin{aligned}
54 < p \leq 60 &\quad \rightarrow \quad A \\
48 < p \leq 54 &\quad \rightarrow \quad B \\
42 < p \leq 48 &\quad \rightarrow \quad C \\
36 < p \leq 42 &\quad \rightarrow \quad D \\
29 < p \leq 36 &\quad \rightarrow \quad E \quad \text{(A pass is guaranteed with the required points for 'E'.)} \\
0 \leq p \leq 29 &\quad \rightarrow \quad F
\end{aligned}
$$

This exam consists of sections **A**, **B**, and **C**. **NB. Use different papers (answer sheets) for different sections.**

A Graded problems

Potential inquiries to be addressed to Atsuto Maki.

A-1 Terminology (4p)

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

a) Bagging

b) Cross validation

c) The LASSO

d) RANSAC

1) A variation of Branch-and-Bound search

2) The latent structure optimization

3) Random strategy for area compression

4) A way to exploit training data for assessing a model

5) An example of ensemble learning

6) A method for evaluating a mixture of models

7) A regularization method that results in feature seletion

8) Robust method to fit a model to data involving outliers

Solution: a-5, b-4, c-7, d-8

A-2 Entropy (6p)

You have booked a flight for tomorrow, but now there are some risks that it might be cancelled due to two factors: typhoon and strike. Your estimate on flight cancellation due to the weather, i.e. typhoon, is 40%. Independently, the probability of cancellation due to strike is 50%.

a) What is the probability that there will be a flight cancellation due to one or both of the factors? (2p)

b) How unpredictable is it that the flight is cancelled, either due to the weather or the strike (or both)? Answer in terms of Entropy, measured in bits. (2p)

c) You realized that you can find out whether there will be a strike or not (which we can assume reliable) on the airport website tonight.

What is the expected information gain from checking it on the website? (2p)

Solution:

a) 1 - (The probability that there will be no flight cancelletion = 0.60 $\times$ 0.50) = 0.70.

b) -0.7 log_2 0.7 - 0.3 log_2 0.3 $\approx$ 0.881 (bits)

c) With proobability 0.5 you will know for sure (Ent = 0) that the flight is canceled. Otherwise, the only uncertainly that remains is the one caused by the weather which has the entropy: -0.6 log_2 0.6 - 0.4 log_2 0.4 $\approx$ 0.971.
Expected gain = 0.881 - (0.5 $\times$ 0 + 0.5 $\times$ 0.971) = 0.396 (bits)

A-3 Bias and Variance (5p)

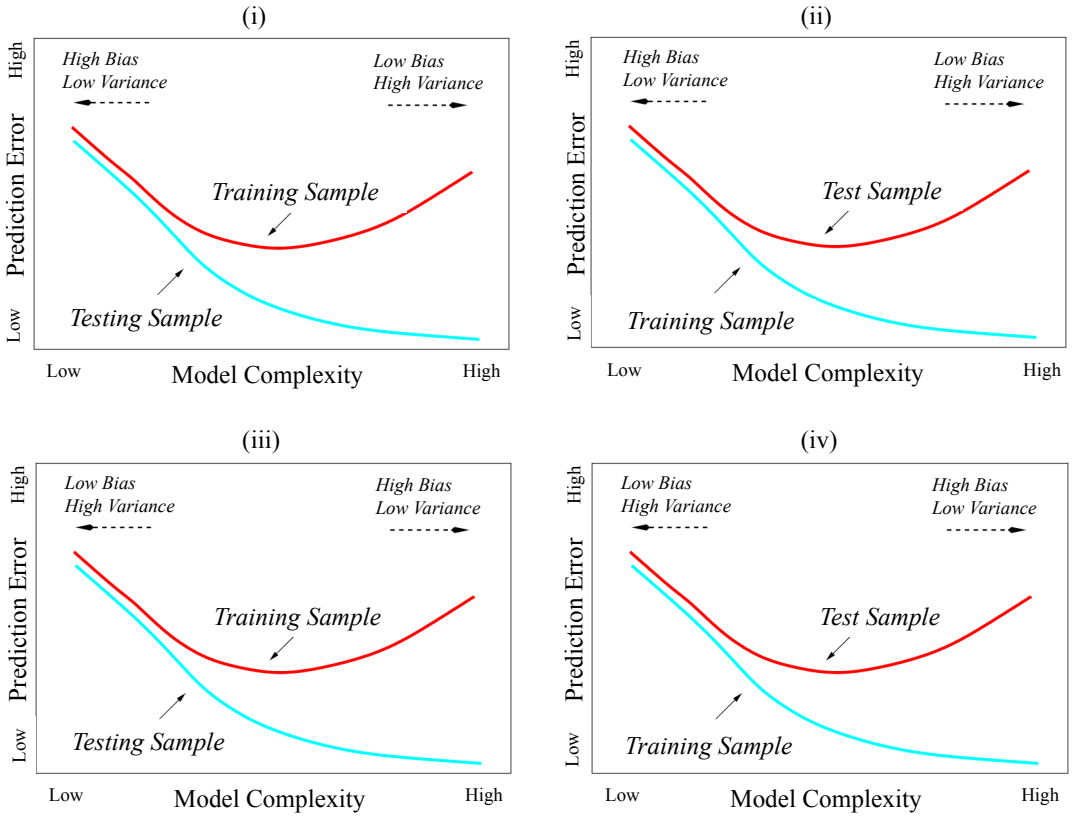


Figure 1. Typical behavior of prediction error plotted against model complexity.

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}. Which one of the four figures most well represents the general situation? (1p)

b) Now, let us consider a model, function $f(\mathbf{x})$ of input vector $\mathbf{x}$, and the following concepts:

$$\begin{aligned} \hat{f}(\mathbf{x}) &: \text{prediction function (= model) estimated with a set of data samples, } \mathcal{D} \\ E_{\mathcal{D}}[\hat{f}(\mathbf{x})] &: \text{the average of models due to different sample set} \end{aligned}$$

Show the bias and variance of the model *in formulae* referring these terms. (2p)

c) Derive that the mean square error (MSE) for estimating $f(\mathbf{x})$ can be decomposed into a two-fold representation consisting of the terms of bias and variance. (2p)

Solution:

a) (ii)

b) Bias: $E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x})$
Variance: $E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])^2]$

c)

$$
\begin{aligned}
& E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] \\
= \; & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})] + E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2] \\
= \; & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])^2 + (E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2 + 2(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])(E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))] \\
= \; & E_{\mathcal{D}}[(\hat{f}(\mathbf{x}) - E_{\mathcal{D}}[\hat{f}(\mathbf{x})])^2] + (E_{\mathcal{D}}[\hat{f}(\mathbf{x})] - f(\mathbf{x}))^2 \\
= \; & Variance + (Bias)^2
\end{aligned}
$$

A-4 Ensemble Methods (4p)

Briefly answer the following questions regarding ensemble methods of classification.

a) What are the two kinds of randomness involved in the design of Random 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? How are they used?

Solution:

a) Randomness (i) in generating bootstrap replicas and (ii) in possible features to use at each node.

b) The update is according to (i) if the sample was misclassified (then the weight is increased, vice-versa), and (ii) the reliability of the weak classifier based on the training error used as the coefficient; the smaller the training error, the greater the reliability.

A-5 Principal Component Analysis (PCA), Subspace Method (3p)

We consider to solve a K-class classification problem with the Subspace Method and for that we compute a subspace $\mathcal{L}^{(j)}$ $(j = 1, ..., K)$ using training data for each class, respectively. That is, given a set of feature vectors (as training data) which belong to a specific class C (i.e. with an identical class label), we perform PCA on them and generated an orthonormal basis $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ which spans a p-dimensional subsapce, $\mathcal{L}$, as the outcome.

Provide an answer to the following questions.

a) Which one among the three is least relevant to PCA?

i. Viewing the input data from a different coordinate system.
ii. Finding errors in data labels.
iii. Exploring the possiblity of dimentionality reduction.

Simply indicate your choice. (1p)

b) Given that we compute $\{\mathbf{u}_1, ..., \mathbf{u}_p\}$ as eigenvectors of the auto-correlation matrix Q based on the training data, how should we choose the eigenvectors in relation to the corresponding eigenvalues of Q? (1p)

c) Given $\mathbf{x}$, we computed its projectiton length on each subspace as $S^{(j)}$ ($j = 1, ..., K$), respectively. For classes with labels $\{l,m,n\}$ among those, we had the following observations:
$S^{(l)}$ was the minimum of all $S^{(j)}$'s,
$S^{(m)}$ was 0.5, and
$S^{(n)}$ was the maximum of all $S^{(j)}$'s.
Which class should $\mathbf{x}$ belong to? Simply choose a class label. (1p)

Solution:

a) ii

b) Choose the eigenvectors corresponding to the largest eigenvalues of Q.

c) n

B Graded problems

Potential inquiries to be addressed to zoom link (B).

B-1 Warm up (4p)

Our dog Shoogee (pictured) loves to hide from us and surprise us when we find her – followed by high pitched squeals of love and tickles. One day she hid from us. We were 99% sure she was in one of the nine hiding spots we know of in the apartment. She chooses any of the spots with equal frequency, so we just started searching at random. We checked all but one of the spots, and still no Shoogee. What's the probability she has found a new hiding spot?

Figure 2. Shoogee (dog) in one of her nine known hiding spots.

Solution: The sample space has three outcomes: A) Shoogee is in the one known spot; B) Shoogee is in any of the other known spots; and C) Shoogee is not in any of the known spots. These outcomes are mutually exclusive, and the union of them is the sure event, i.e., $P(A) + P(B) + P(C) = 1$. From our prior belief, $P(A) + P(B) = p = 0.99$. This means $P(C) = 1 - p = 0.01$. Since Shoogee chooses any of the $N = 9$ hiding spots with the same frequency, $P(A) = p/N$.

We have observed $\neg B$, and would like to compute $P(C|\neg B) = 1 - P(A|\neg B)$, from the fact that $\neg B$ makes the sample space be the union of A and C. By Bayes' Theorem, $P(A|\neg B) = P(\neg B|A)P(A)/P(\neg B)$. Now, $P(\neg B|A) = 1$ because A automatically implies $\neg B$. Hence $P(A|\neg B) = P(A)/P(\neg B) = (p/N)/P(\neg B)$. The denominator is $P(\neg B) = P(A) + P(C) = p/N + (1 - p)$. Putting it all together

$$P(C|\neg B) = 1 - \frac{p/N}{p/N + (1-p)} = \frac{N(1-p)}{p + (1-p)N} = \frac{N}{\frac{p}{1-p} + N} = \frac{9}{99 + 9} = \frac{1}{12} \tag{1}$$

B-2 Maximum likelihood estimation (4p)

Consider the data in Table 1. Assume all observations (walks) are independent. Assume the number of poopoos emitted by Shoogee Nulnul is distributed Poisson with parameter λ_S. Assume the number of poopoos emitted by Max the Tax conditioned on the number of poopoos emitted by Shoogee Nulnul s is distributed Poisson with parameter $\lambda_M + s$. Find the maximum likelihood estimates of λ_S and λ_M. Show your work. (You may use a calculator.)

	poopoos	
walk	Shoogee Nulnul	Max the Tax
1	0	1
2	1	2
3	1	1
4	0	1
5	1	2

Table 1. The number of poopoos emitted by two doggos on five walks together.

Solution: The dataset is $\mathcal{D} = \{(s, m)_w : w \in [1, 5]\}$. Define the random variables S and M as the number of poopoos emitted by Shoogee Nulnul and Max the Tax, respectively. Assuming all observations are independent, the likelihood of the dataset is

$$L(\mathcal{D}) = \prod_{w=1}^{5} P[S = s_w, M = m_w] = \prod_{w=1}^{5} P[S = s_w]P[M = m_w|S = s_w] \tag{2}$$

by the definition of conditional probability. The log-likelihood is

$$\log L(\mathcal{D}) = \sum_{w=1}^{5} (\log P[S = s_w] + \log P[M = m_w|S = s_w]) \tag{3}$$

We know these distributions from the wording of the problem:

$$P[S = s] = \frac{\lambda_S^s e^{-\lambda_S}}{s!} \tag{4}$$

$$P[M = m|S = s] = \frac{(\lambda_M + s)^m e^{-(\lambda_M + s)}}{m!} \tag{5}$$

Plugging these into the above expression gives:

$$\log L(\mathcal{D}) = \sum_{w=1}^{5} (\log \frac{\lambda_S^{s_w} e^{-\lambda_S}}{s_w!} + \log \frac{(\lambda_M + s_w)^{m_w} e^{-(\lambda_M + s_w)}}{m_w!}) \tag{6}$$

$$= \sum_{w=1}^{5} s_w \log \lambda_S - \lambda_S - \log s_w! + m_w \log(\lambda_M + s_w) - (\lambda_M + s_w) - \log m_w! \tag{7}$$

We want to maximize this for each parameter, so taking the derivate and setting to zero:

$$\frac{\partial}{\partial \lambda_S} \log P(D) = \sum_{w=1}^{5} (s_w \frac{1}{\lambda_S} - 1) = 0 \tag{8}$$

$$\Rightarrow \lambda_S = \frac{1}{5} \sum_{w=1}^{5} s_w = 0.6 \tag{9}$$

$$\frac{\partial}{\partial \lambda_M} \log P(D) = \sum_{w=1}^{5} (m_w \frac{1}{\lambda_M + s_w} - 1) = 0 \tag{10}$$

$$\Rightarrow 5 = \frac{2}{\lambda_M} + \frac{4}{\lambda_M + 1} \tag{11}$$

$$\Rightarrow 5\lambda_M(\lambda_M + 1) = 6\lambda_M + 2 \tag{12}$$

$$\Rightarrow 5\lambda_M^2 - \lambda_M - 2 = 0 \tag{13}$$

$$\Rightarrow \lambda_M = \frac{1 \pm \sqrt{1 - 4(5)(-2)}}{10} = \frac{1 \pm \sqrt{41}}{10} = \frac{1 + \sqrt{41}}{10} \approx 0.74 \tag{14}$$

since the parameter $\lambda_M > 0$.

B-3 Maximum a posteriori classification (4p)

Consider the data in Table 1. Assume all observations (walks) are independent. Assume the number of poopoos emitted by Shoogee Nulnul is distributed Poisson with parameter $\lambda_S = 0.60$. Assume the number of poopoos emitted by Max the Tax conditioned on the number of poopoos emitted by Shoogee Nulnul s is distributed Poisson with parameter $\lambda_M = 0.74$. On a sixth walk together, one of the doggos emitted 1 poopoos. Which dog is the most likely to have done this according to a maximum a posteriori classifier? Show your work. (You may use a calculator.)

Solution: Denote P as the number of poopoos, and $d \in \{\text{Shoogee}, \text{Max}\}$ be the doggos. The maximum a posteriori classifier is defined

$$d_{\text{MAP}} = \arg \max_{d \in \{\text{Shoogee,Max}\}} P(P = p|D = d)P(D = d) \tag{15}$$

$$= \arg \max_{d \in \{\text{Shoogee,Max}\}} P(P = p|D = d). \tag{16}$$

since the prior $P(D = d) = 0.5$ because both doggos are on the walk. We just need to find an expression for $P(P = p|D = d)$. For Shoogee Nulnul,

$$P[P = p|D = \text{Shoogee}] = \frac{\lambda_S^p e^{-\lambda_S}}{p!} \Rightarrow P[P = p|D = \text{Shoogee}] = \frac{0.6^p e^{-0.6}}{p!}. \tag{17}$$

For Max the Tax, we want to find

$$P[P = p|D = \text{Max}] = \sum_{s=0}^{\infty} P[M = p, S = s] = \sum_{s=0}^{\infty} P[M = p|S = s]P[S = s] \tag{18}$$

where M is the number of poopoos by Max the Tax and S is the number of poopoos by Shoogee Nulnul. Finally, plugging in what we know:

$$P[P = p|D = \text{Max}] = \sum_{s=0}^{\infty} \frac{(0.74 + s)^p e^{-(0.74+s)}}{p!} \frac{0.6^s e^{-0.6}}{s!} \tag{19}$$

When $P = 1$ then

$$P[P = 1|D = \text{Shoogee}] = 0.6e^{-0.6} \approx 0.33 \tag{20}$$

$$P[P = 1|D = \text{Max}] = \sum_{s=0}^{\infty} (0.74 + s)e^{-(0.74+s)} \frac{0.6^s e^{-0.6}}{s!} \approx 0.31. \tag{21}$$

Hence, the likely culprit is Shoogee Nulnul.

C Graded problems

Potential inquiries to be addressed to Jörg Conradt.

C-1 Multiple-Choice: Support Vector Machine (1p)

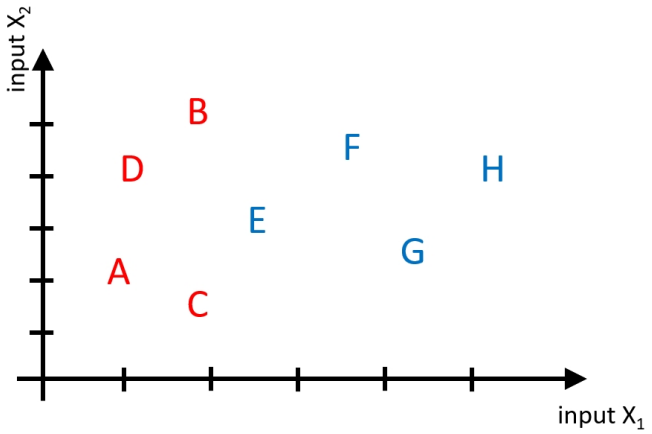
Do not justify your answer. Instead, select exactly **one** option of (1.), (2.), or (3.).
Complete the following sentence: Out of all hyperplanes which solve a classification problem, the one with widest margin will probably ...

1. ... generalize best.
2. ... compute fastest.
3. ... have the smallest number of parameters.

Solution: 1

C-2 Support Vector Classification (3p)

The following diagram shows a small data set consisting of four RED samples (A, B, C, D) and four BLUE samples (E, F, G, H). This data set can be linearly separated.

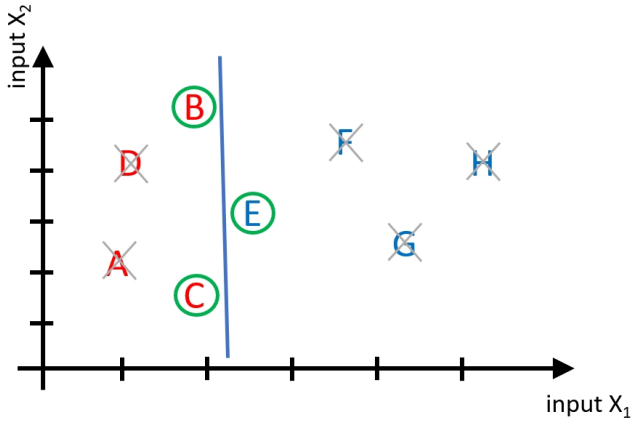


a) We use a linear support vector machine (SVM) without kernel function to correctly separate the RED (A-D) and the BLUE (E-H) class. Which of the data points (A-H) will the support vectors machine use to separate the two classes? Name the point(s) (no explanation needed). (1p)

b) Assume someone suggests using a non-linear kernel for the SVM classification of the above data set (A-H). Give one argument **in favor of** and one argument **against** using non-linear SVM classification for such a data set. **USE KEYWORDS!** (2p)

Solution:

a) The blue line shows the linear decision boundary between the two classes.



The data points in green circles are support vectors for the linear decision boundary**B**, **C**, and **E**. (Reason: if any of those data points change, the decision boundary changes) - note: reason not needed.

b) at least one of the following (or similar) **in favor** (1p):

+) The decision boundary margin might get wider with a non-linear kernel.

+) The same learning approach is likely to work for additional (possibly more complex) data.

at least one of the following (or similar) **against** (1p):

+) Computational complexity.

+) programming / computing complexity.

C-3 Multiple-Choice: Artificial Neural Networks (1p)

Do not justify your answer. Instead, select exactly **one** option of (1.), (2.), or (3.).
Error-Backpropagation-Training for Neuronal Networks requires the following:

1. a single input and a single output to the network.
2. labeled training data and differentiable neuron activation functions.
3. Gaussian distributed training data and a fast computer.

Solution: 2

C-4 Artificial Neural Networks (3p)

Consider the training data in the table, where $+$ means a positive sample and $-$ a negative.

x_1	x_2	Class
8	8	$-$
8	-4	$-$
4	0	$+$
0	4	$+$
-6	-6	$-$
-6	8	$-$

a) What is the minimum number of *layers* needed for an artificial neural network to correctly classify all these points? Motivate your answer **IN KEYWORDS**. (2p)

b) How many input nodes and how many output nodes does your neuronal network need to address this problems? (1p)

Solution:

a) Two layers are 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 and one output node (given by the data table).