UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Midterm Exam
Date: 26 September 2014
Time: 11.00-12.30

Number of pages: 4 (including front page)
Number of questions: 6
Maximum number of points to earn: 60
At each question is indicated how many points it is worth.

BEFORE YOU START

- Please **wait** until you are instructed to open the booklet.
- Check if your version of the exam is complete.
- Write down **your name, student ID number**, and if applicable the **version number** on **each sheet** that you hand in. Also **number the pages**.
- Your **mobile phone** has to be switched off and in the coat or bag. Your **coat and bag** must be under your table.
- **Tools allowed**: A single A4 cheat-sheet.

PRACTICAL MATTERS

- The first 30 minutes and the last 15 minutes you are not allowed to leave the room, not even to visit the toilet.
- You are obliged to identify yourself at the request of the examiner (or his representative) with a proof of your enrollment or a valid ID.
- During the examination it is not permitted to visit the toilet, unless the proctor gives permission to do so.
- 15 minutes before the end, you will be warned that the time to hand in is approaching.
- If applicable, please fill out the evaluation form at the end of the exam.

Good luck!

Questions

1. In Amsterdam, scooters are allowed to ride on the bike paths if they have a blue license plate, but are not allowed on the roads. The reverse is true for scooters with yellow license plates. The city estimates that at any given time, anywhere in the city, that a scooter on a bike path is yellow 1% of the time (i.e. the vast majority of the scooters stick to where they belong).

   One evening there is a hit-and-run accident between a scooter and a cyclist on a bike path. A witness tells police that the scooter had a yellow license plate. The police want to assess the reliability of the witness by testing him with different scooters under the same conditions the evening of the accident. The witness correctly identifies the colour of a license plate 9/10 times. In other words, if the police test the witness with a blue bike, the witness will claim they saw blue 9 of 10 tests with blue; the same is true for testing and claiming a yellow bike.

   We introduce a discrete random variable C for license plate colour that can take values y or b (yellow or blue). We are interested in the probability of the colour of a scooter's license plate *on the bike path.* We also introduce a discrete random variable W for the color that a witness claims to see that can take on values y and b (yellow or blue).

   Given this information, answer the following questions:

   (a) What is $P(C = b)$ and $P(C = y)$ on a bike path? **/2**

   (b) What is the probability that the accident was caused by a yellow licensed scooter, if the witness claims it was yellow? I.e. what is $P(C = y|W = y)$? **/3**

   (c) If there was no witness, that would be the probability that the accident was caused by a blue plate? **/1**

2. A hospital has a computer program that can classify X-ray images. Given an image, the program will output the probability of a tumour in the image, i.e. $P(t = 1|\mathbf{x})$. The classifier has been trained on a large data set and it has been tested and performs well. The hospital uses the X-ray as a cheap first step in the diagnosis of tumours. Possible actions based on the X-ray classification are a_0 (discharge) and a_1 (take a second test, an expensive MRI). Given this information, answer the following questions:

   (a) For classification in general, what is the optimal Bayesian decision rule for choosing class k given K posterior probabilities $P(\mathcal{C}_k|\mathbf{x})$? **/2**

   (b) What is the optimal Bayesian decision rule in this example? **/2**

   (c) What loss does this rule minimize? **/1**

   (d) In this example, however, there is one (non-zero) loss associated with the each action. There is a loss of $L_0 = 1000$ if the patient is discharged and there is really a tumour and there is a loss of $L_1 = 1$ if the patient does not have a tumour, but the MRI test is performed. What is the decision rule with the loss-adjusted conditional probabilities? **/3**

3. In the following we assume that we have a conditional model $p(t|\phi, \mathbf{w})$, with t a real-valued target, ϕ an M-dimensional vector of basis functions. We are free to choose M. Assume that as M increases, the complexity of the model increases. Assume we have a training set $\mathcal{D}_{train}$ and a test set $\mathcal{D}_{test}$.

   (a) Draw a graph, indicating qualitatively how the training and test errors change as a function of M for a typical learning problem. /2

   (b) Indicate on the graph where i) overfitting is occurring, ii) underfitting is occurring, and iii) the best choice for M. /3

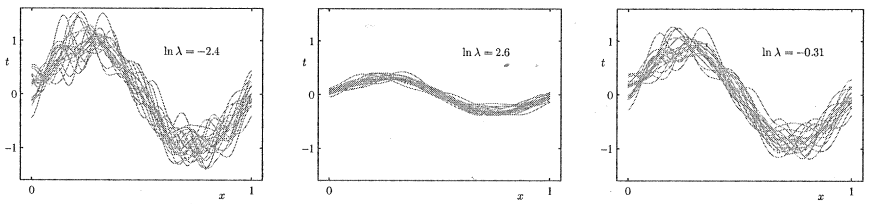
   (c) What is overfitting and why does it occur? /2

   (d) What is underfitting and why does it occur? /2

   (e) Assume we do not have $\mathcal{D}_{test}$. What can we do to detect overfitting? /2

   (f) Briefly describe two approaches we can use to avoid overfitting. /2

4. There are three figures below corresponding to three different weight penalties λ. The red lines are the predictions based on the penalized least-squares regression model. Each line is a prediction learned from a randomly generated data set.



(A) (B) (C)

   (a) Write down the letter (A/B/C) associated with i) high-bias ii) low-bias iii) high-variance. /3

   (b) Which models (high/low-bias and high/low variance) tend to overfit? Underfit? /2

   (c) What are two approaches to decrease the variance of a predictive model? /2

5. Assume a classification problem with two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. Given $p(\mathcal{C}_1) = \pi_1$, $p(x|\mathcal{C}_1) = \mathcal{N}(x|u_1, \sigma^2)$, $p(x|\mathcal{C}_2) = \mathcal{N}(x|u_2, \sigma^2)$, answer these questions:

   (a) Find a general expression for $p(\mathcal{C}_2|x)$ using $p(\mathcal{C}_1)$, $p(x|\mathcal{C}_1)$, $p(x|\mathcal{C}_2)$. /1

   (b) What is the inequality that describes the conditions for classifying x as $\mathcal{C}_1$? /1

   (c) Starting with this inequality, solve for the condition on x that will classify x as $\mathcal{C}_1$, using the specific forms π_1, $\mathcal{N}(x|u_1, \sigma^2)$, $\mathcal{N}(x|u_2, \sigma^2)$? Note: $\mathcal{N}(x|u_1, \sigma^2) = (2\pi)^{-1/2}\sigma^{-1}\exp(-0.5(x-\mu)^2/\sigma^2)$. /5

   (d) Using your solution above, assume $\pi_1 = 0.5$ and $\mu_1 = 2$ and $\mu_2 = 1$. What is the decision boundary $\hat{x}$? /2

   (e) What is the effect on the decision boundary if $\pi_1 << \pi_2$? Why does this effect make sense? /3

6. In this question we will consider Bayesian linear regression. Let $\mathbf{w}$ be a vector of regression weights, $\boldsymbol{\phi}_\star$ a test vector, $t_\star$ the prediction at the test vector, and training data set $\mathcal{D} = \{\boldsymbol{\Phi}, \mathbf{t}\}$.

   (a) Consider first the general case, where we do not assume a specific distribution for the posterior or posterior predictive distribution. Assume we have the posterior distribution of $\mathbf{w}$, i.e. $p(\mathbf{w}|\mathcal{D})$ and a likelihood for the test prediction $t_\star$ given a test vector $\boldsymbol{\phi}_\star$, i.e. $p(t_\star|\boldsymbol{\phi}_\star, \mathbf{w})$. What is the exact form of the posterior predictive distribution? Indicate the two rules of probability are required to solve for this distribution. /3

   (b) Now consider the specific case (which was covered in Bishop and your homework) of a Gaussian posterior for $\mathbf{w}$ (the distribution and its parameters are described below). Convert the posterior $\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ and its parameters into *scalar* form ($\mathbf{m}_N \rightarrow m_N$ and $\mathbf{S}_N \rightarrow \sigma_N^2$). I.e. derive the simplified forms m_N and σ_N^2. /2

posterior of $\mathbf{w}$:

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$
$$\mathbf{S}_N = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}$$
$$\mathbf{m}_N = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\beta\boldsymbol{\Phi}^T\mathbf{t}$$

   (c) Using the scalar form, compute limit of the posterior variance σ_N^2 as $N \rightarrow \infty$. /2

   (d) How do you interpret this limit? /2

   (e) Continuing with the specific Gaussian case, derive the parameters for the posterior predictive $\mathcal{N}(t_\star|y_\star, 1/\beta_\star)$ for the *scalar* form (using the same symbols). The non-scalar distribution is given below. /2

posterior predictive of $t_\star$:

$$p(t_\star|\boldsymbol{\phi}_\star, \mathbf{t}, \boldsymbol{\Phi}, \alpha, \beta) = \mathcal{N}(t_\star|y_\star, 1/\beta_\star)$$
$$y_\star = \boldsymbol{\phi}_\star^T\mathbf{m}_N$$
$$\beta_\star = \left(1/\beta + \boldsymbol{\phi}_\star^T\mathbf{S}_N\boldsymbol{\phi}_\star\right)^{-1}$$

   (f) Using the scalar form, compute limit of the posterior predictive precision $\beta_\star$ as $N \rightarrow \infty$. How do you interpret this limit? /2

   (g) What two uncertainties does $\beta_\star$ encode? /2