UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Final Exam
Date: 21 October 2014
Time: 13.00-15.00

Number of pages: 4 (including front page)
Number of questions: 3
Maximum number of points to earn: 85
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!

1 Principal Component Analysis

Suppose we have a data set $\{\mathbf{x}_n\}_{n=1}^N$ of D-dimensional vectors that have been *centered* such that $\sum_{n=1}^N x_{nd} = 0 \; \forall d$. **/25**

(a) Provide an expression for the sample covariance $\mathbf{S}$ of $\{\mathbf{x}_n\}_{n=1}^N$. **/2**

(b) Assume we perform a **complete** eigenvalue decomposition $\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^T$.

(i) What is the dimensionality of $\mathbf{U}$ and Λ? **/1**

(ii) What is entry (i, j) of Λ when $i \neq j$ and $i = j$? **/1**

(iii) Let $\mathbf{u}_i$ be the ith column of $\mathbf{U}$. What are the values of the entries of the vector $\mathbf{u}_i^T\mathbf{U}$? **/1**

(c) Write down an expression for $\mathbf{x}_n$ in terms of D principal components (i.e. the complete set or full basis) **and** $\tilde{\mathbf{x}}_n$, an approximation of $\mathbf{x}_n$ based on the **first** K principal components. Remember, $\{\mathbf{x}_n\}_{n=1}^N$ have been centered. **/5**

(d) Assuming the representations from (c),

(i) Find an expression for the error $E = \frac{1}{N}\sum_{n=1}^N ||\mathbf{x}_n - \tilde{\mathbf{x}}_n||^2$, in terms of eigenvalues Λ_{ii}, $i > K$. **/7**

(ii) Are these the *largest* or *smallest* eigenvalues of the complete set? **/1**

(e) Briefly describe what PCA does geometrically. Use a drawing to illustrate your description. **/4**

(f) Briefly describe sphering. Include an additional drawing (from (e)) showing the final stages of the sphering operation. **/3**

2 Outlier detection using ν-SVM

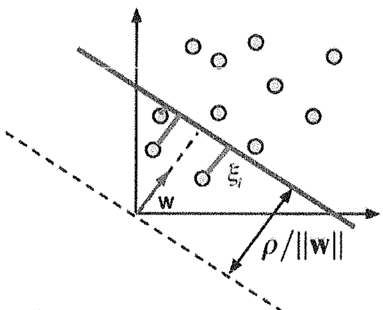
/30

We are given the following dataset: $\{\mathbf{x}_n\}$, $i = 1..N$, where each $\mathbf{x}_n \in \mathbb{R}^D$. We are also given a collection of feature functions $\{\phi_a(\cdot)\}$, $a = 1..A$. Now consider the following optimization problem for outlier detection,

$$\min_{\mathbf{w},\{\xi_n\},\rho} \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{\nu N}\sum_{n=1}^{N} \xi_n$$

$$\text{subject to: } \mathbf{w}^T\phi_n(\mathbf{x}_n) \geq \rho - \xi_n, \quad \xi_n \geq 0 \quad \forall n$$

where $\mathbf{w}$ is a weight vector, ρ is a offset, ξ_n are slack variables. The parameter $\nu \in [0,1]$ represents the fraction of training vectors that are treated as outliers. The expression $\frac{1}{\nu N}$ plays a similar role to C in regular SVMs. The figure below illustrates the geometry of the problem for a linear kernel function: the goal is to separate ν% of the training data from the rest of the data (ν is usually small, for example 0.01 or 0.05). The hyperplane separates the data, treating the data as 2 classes, $y(\mathbf{x}) = +1$ for *normal* data and $y(\mathbf{x}) = -1$ as *outliers* (which are closest to the origin).



(a) Provide an expression for the primal Lagrangian. Use Lagrange multipliers $\{\alpha_n\}$ and $\{\beta_n\}$. /5

(b) Write down and/or solve for all the KKT conditions. /10

(c) Use these conditions to write down the dual Lagrangian. Make sure to include the dual constraints. /5

(d) Assume we solve the dual program which gives $\{\alpha_n^\star\}_{n=1}^N$. How do we determine the support vectors? What value of ξ to they have? /5

(e) The solution does not provide an explicit value for ρ.

(i) Reason about the KKT conditions, in particular the complementary slackness conditions, to determine a precise expression for ρ based on a single training example ϕ_n. /3

(i) How can we improve the numerical stability of this solution by using the entire training set? What is the expression for ρ in this case? /2

(f) Assume we have a test vector $\mathbf{x}_\star$. We want to test whether it is a normal vector or an outlier. What is the prediction $y(\mathbf{x}_\star)$? The expression may only involve kernel evaluations (instead of feature evaluations). /3

(g) Describe the solution of the dual program when $\nu = 1$? In particular, what happens to the values of the support vectors? /2

3 Mixture Models

/30

In this question we consider an extension of the linear regression model to a **mixture of linear regression models**. The predictive distribution (likelihood) for output t_n, conditioned on its input vector $\mathbf{x}_n$ is:

$$p(t_n|\mathbf{x}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(t_n|\mathbf{w}_k^T\mathbf{x}_n, \beta^{-1}\right)$$

where k is the index of the mixture component, $\mathbf{w}_k$ are D-dimensional regression weights for component k, and β is a global precision parameter. Note: $\mathcal{N}(x|u, \sigma^2) = (2\pi)^{-1/2}\sigma^{-1}\exp(-0.5(x-\mu)^2/\sigma^2)$. To answer the following questions assume we are given a dataset of input vectors $\mathbf{x}_n \in \mathbb{R}^D$, $n = 1..N$ and output scalars $t_n \in \mathbb{R}$. When you answer the questions below, ensure that the constraint $\sum_k \pi_k = 1$ is satisfied. We can define the expression for responsibility r_{nk} as

$$r_{nk} = \frac{\pi_k \mathcal{N}\left(t_n|\mathbf{w}_k^T\mathbf{x}_n, \beta^{-1}\right)}{\sum_j \pi_j \mathcal{N}\left(t_n|\mathbf{w}_j^T\mathbf{x}_n, \beta^{-1}\right)}$$

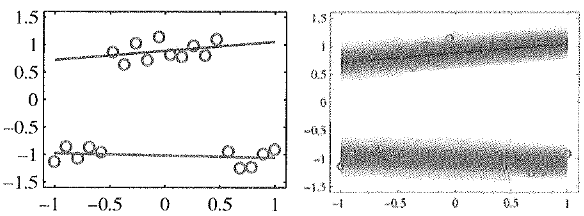


Figure 1: A mixture of 2 linear regression models. The x-axis represents the input ($\mathbf{x}$) and the y-axis represents the output (t). The expected predictions by cluster k are indicated by coloured lines (left) along with the full predictive densities (right).

(a) Write down the log-likelihood for the data set in terms of $\{t_n, \mathbf{x}_n\}$, $\{\pi_k\}$, $\{\mathbf{w}_k\}$, and β. /5

(b) Find the expression for π_k that maximizes the log-likelihood. When solving for π_k, you should construct and identify $\{r_{nk}\}$ and assume it is fixed. /5

(c) Find the expression for $\mathbf{w}_k$ that maximizes the log-likelihood. Hint: use $\mathbf{R}_k = \text{diag}(r_{nk})$ to simplify the expression (i.e. $\mathbf{R}_k$ is a N by N diagonal matrix with entry n, n equal to r_{nk}). /5

(d) Find the expression for β that maximizes the log-likelihood. /5

(e) Write down an iterative algorithm using the above update equations (similar to the ones derived in class for the Mixture of Gaussians); include initialization and convergence check steps. /5

(f) This model gives significant predictive mass to regions without data (see Figure 1, right). Explain why this occurs and how replacing π_k with the function $\pi_k(\mathbf{x}_n)$ can improve the model. What functional form or model can we use for $\pi_k(\mathbf{x}_n)$? /5