UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Resit Exam
Date: January 8, 2018
Time: 18:00-21:00

Number of pages: 10 (including front page)
Number of questions: 5
Maximum number of points to earn: 46
At each question the number of points you can earn is indicated.

BEFORE YOU START

- As soon as you receive your exam you may start.
- 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**: 1 handwritten double-sided A4-size cheat sheet, pen.
- Multiple choice answers must be indicated on the exam booklet.

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.
- Please fill out the evaluation form at the end of the exam.

Good luck!

1 Multiple Choice Questions

/12

For the evaluation of each question note: several answers might be correct and at least one is correct. You are granted one point if every correct answer is 'marked' **and** every incorrect answer is 'not marked'. For each mistake a 1/2 point is deducted, with the minimum possible number of points per question equal to 0. A box counts as 'marked' if a clearly visible symbol is written in there or if the box is blackened out. In the case you want to change an already marked box write 'not marked' next to the box.

1. Which of the following problems or algorithms correspond to unsupervised learning: **/1**

   [X] Anomaly detection with support vector machines.

   [X] Dimensionality reduction.

   [X] K-means clustering.

   [ ] The Perceptron algorithm

2. Which of the following equations are correct? **/1**

   [ ] $p(x) = \int p(x, y)p(y)dy$.

   [ ] The probability that a continuous random variable x takes on a value in the interval (a, b) with $b > a$ is given by $p(x \in (a, b)) = \int_a^{a+b} p(x)dx$.

   [X] $p(x, y) = p(x|y)p(y)$.

   [X] $p(x, y) = p(y|x)p(x)$.

3. Consider regularized linear regression with the error function $E(\mathbf{w}, \lambda) = \frac{1}{2N} \sum_{n=1}^{N} \left(\boldsymbol{\phi}(\mathbf{x}_n)^T \mathbf{w} - t_n\right)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$. You want to find the optimal parameter for the regularization penalty $\lambda \in \{0.1,\ 0.01,\ 0.001\}$. You split your dataset consisting of N_{tot} datapoints into a training set, a validation set and a test set. You obtain the following validation and training errors $E_{\text{val}}(\mathbf{w}, \lambda)$, and $E_{\text{train}}(\mathbf{w}, \lambda)$:

	$E_{\text{train}}(\mathbf{w}, \lambda)$	$E_{\text{val}}(\mathbf{w}, \lambda)$
$\lambda_1 = 0.1$	1.72	2.1
$\lambda_2 = 0.01$	0.63	0.85
$\lambda_3 = 0.001$	0.45	1.34

   Which of the following statements are correct? **/1**

   [ ] If we retrain our model with a larger dataset, we do not need to re-estimate the optimal value of λ.

   [ ] λ_1: overfitting, λ_2: best fit, λ_3: underfitting.

   [X] λ_1: underfitting, λ_2: best fit, λ_3: overfitting.

   [ ] λ_1: underfitting, λ_2: best fit, λ_3: underfitting.

4. You are given a dataset $\mathcal{D} = \{x_n\}_{n=1}^N$. The data is normally distributed with $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over μ : $\mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. Indicate which of the following statements are correct: **/1**

   - [X] When N (the number of datapoints in D) becomes larger, the posterior distribution over μ will become narrower.
   - [ ] When N (the number of datapoints in D) becomes larger, the prior distribution over μ will become wider.
   - [ ] The estimate of the posterior distribution over μ is insensitive to the choice of the prior for finite N.
   - [X] When σ_0^2 is small, the prior constrains the maximum a posteriori estimate of μ strongly.

5. Which of the following statements about classification are correct? **/1**

   - [ ] Logistic regression is a probabilistic generative model.
   - [X] For a Naive Bayes classifier with feature vectors $\mathbf{x} = (x_1, x_2, ..., x_D)$ and K classes $\{\mathcal{C}_1, ..., \mathcal{C}_K\}$, the Naive Bayes assumption is equivalent to assuming $p(\mathbf{x}|\mathcal{C}_k) = p(x_1, x_2, ..., x_D|\mathcal{C}_k) = \prod_{i=1}^D p(x_i|\mathcal{C}_k)$ for $k = 1, ..., K$.
   - [X] The perceptron algorithm is a linear discriminant model for classification.
   - [ ] In probabilistic discriminative models, the prior probability of class C: $p(C)$ is modeled.

6. Consider a neural network with two layers, and 10 hidden units in the hidden layer. Which of the following statements are correct? **/1**

   - [ ] For a regression task with a target $t > 0$, applying the activation function $f(x) = x$ to the output unit ensures that the model outputs numbers in the correct range.
   - [ ] For classification with K mutually exclusive classes we need $2K$ output units with the softmax activation function applied to the output units.
   - [ ] For classification with $K = 4$ *not* mutually exclusive classes, a suitable activation function for the output units is the softmax activation function.
   - [X] For regression with targets $-1 < t < 1$, a suitable activation function for the output unit is $f(x) = \tanh(x)$.

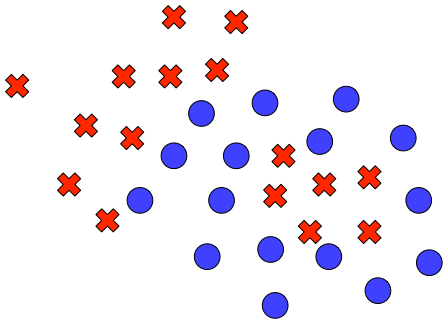
7. Which of the following statements about training a neural network with stochastic gradient descent (SGD) are correct? **/1**

- [X] SGD can be performed by sampling single data points, or with minibatches.
- [ ] In forward propagation we propagate errors forward through the network in order to evaluate derivatives.
- [X] Full batch gradient descent is more sensitive to get stuck in local minima than stochastic gradient descent.
- [ ] If the learning rate in SGD is too low the algorithm can keep oscillating around a local minimum without converging.

8. Which of the following statements about the EM algorithm for Gaussian Mixture Models are correct? **/1**

- [ ] The updates for the covariance matrices in the M step ensure that the clusters will always have a spherical shape.
- [ ] In the EM algorithm, the M step corresponds to computing the responsibilities with all other parameters fixed.
- [ ] The E step consists of updating the means, the covariances and the cluster mixture components.
- [X] After an update of the parameters using an E step, followed by an M step, the log likelihood function is guaranteed to have increased.

9. Which of the following classifiers can learn the decision boundary for the dataset shown in the figure below? The blue circles belong to one class, and the red crosses belong to the other class.



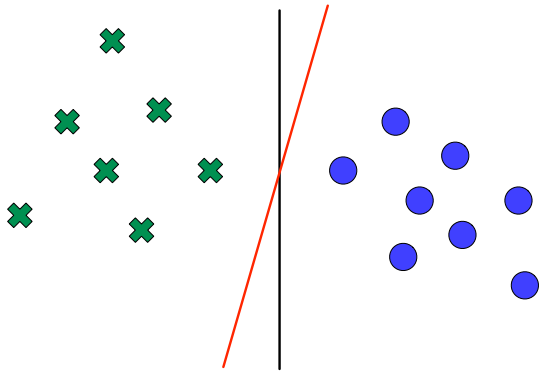
/1

- [X] Neural networks.
- [ ] A perceptron algorithm with linear features.
- [ ] Linear Discriminant Analysis with linear features.
- [ ] Support Vector Machines with a linear kernel.

10. Consider the following two-class SVM optimization problem with data set $\mathcal{D} = \{\mathbf{x}_n, t_n\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^2$ and $t_n \in \{-1, 1\}$:

$$\begin{aligned} \text{minimize}_{\{\mathbf{w},\xi_n,b\}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & t_n(\mathbf{w}^T\mathbf{x}_n - b) \geq 1 - \xi_n \quad \text{for all } n \\ & \xi_n \geq 0 \quad \text{for all } n \end{aligned}$$

The data set is depicted in the figure below. The 8 blue circles correspond to data points with labels $t_n = +1$, and the 7 green crosses represent the data points with labels $t_n = -1$.



Which of the following statements are correct? **/1**

- [X] For $C \to \infty$ the black solid line is the correct decision boundary.
- [ ] For $C \to \infty$ the margin width corresponding to the red solid decision boundary is larger than the margin width corresponding to the black solid decision boundary.
- [X] For $C \to \infty$ the number of support vectors is equal to 2.
- [ ] For $C = 1$ all datapoints will become support vectors.

11. Which of the following statements about Gaussian Processes are correct? Consider a dataset of N datapoints $\{\mathbf{x}_n, t_n\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^M$. **/1**

- [ ] The Gaussian Process model is a parametric model.
- [ ] Consider a Gaussian Process model for regression that has been trained on this dataset. The predictive distribution for a new target t_{N+1} with input vector $\mathbf{x}_{N+1}$, is a Gaussian whose mean depends on $\mathbf{x}_{N+1}$, but the standard deviation does not.
- [X] When $M >> N$, regression with Gaussian Processes is computationally more efficient as compared to regression with a fixed set of basis functions.
- [X] Gaussian Processes can consider covariance functions that can only be expressed in terms of an infinite number of basis functions.

12. Which of the following statements are correct? **/1**

- [X] Consider two independent, normally distributed random variables X and Y, and $\alpha, \beta \in \mathbb{R}$. Then, the random variable $Z = \alpha X + \beta Y$ is also normally distributed.
- [X] The Laplace approximation aims to find a Gaussian approximation to a probability density defined over a set of continuous variables.
- [ ] When applying a Laplace approximation to a multi-modal distribution, the resulting approximation will also be multi-modal.
- [ ] In Gaussian Processes for regression we assume the targets t_n are generated by $t_n = y_n + \varepsilon$, where ε is the same for each datapoint, and is ε sampled from $\mathcal{N}(\varepsilon|0, \beta^{-1})$.

Grading instructions

The solutions given below, with the corresponding distribution of points, serve as a guideline. If some intermediate steps are left implicit by the student, while still clearly following a derivation, points will not be deducted. The total number of possible points is 46, meaning that the final grade is computed as $10 \times \frac{\#\text{points}}{46}$.

General remarks

Start with the questions that you think are easiest. If you get stuck at one subquestion, don't stop but try to solve the next ones, they are not all dependent on each other!

2 L^1-regularized Logistic Regression for K classes

/6

Consider logistic regression for K classes with N training vectors $\{\mathbf{x}_n\}_{n=1}^N$, each of which is mapped to a different feature vector $\boldsymbol{\phi}(\mathbf{x}_n) = \boldsymbol{\phi}_n = (\phi_0(\mathbf{x}_n), \phi_1(\mathbf{x}_n), ..., \phi_{M-1}(\mathbf{x}_n))^T$ using basis functions $\phi_j(\mathbf{x})$ with $j = 0, ..., M-1$, and $\phi_0(\mathbf{x}) = 1$. Each vector $\mathbf{x}_n$ has a corresponding target vector $\mathbf{t}_n$ of size K: $\mathbf{t}_n = (t_{n1}, t_{n2}, ..., t_{nK})^T$, where $t_{nk} = 1$ if $\mathbf{x}_n \in \mathcal{C}_k$, and $t_{nj} = 0$ for all $j \neq k$. The input data can be collected in a matrix $\mathbf{X}$ with row n given by $\mathbf{x}_n^T$, and the targets can be collected in a target matrix $\mathbf{T}$, with row n equal to $\mathbf{t}_n^T$. The feature vectors can also be collected in a matrix $\boldsymbol{\Phi}$ such that the n-th row of $\boldsymbol{\Phi}$ contains $\boldsymbol{\phi}_n^T$:

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \dots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \dots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Assuming i.i.d. data, the posterior class probabilities are modeled by

$$p(\mathcal{C}_k|\boldsymbol{\phi}(\mathbf{x}), \mathbf{w}_1, ..., \mathbf{w}_K) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k(\boldsymbol{\phi}))}{\sum_{j=1}^K \exp(a_j(\boldsymbol{\phi}))},$$

where $a_k(\boldsymbol{\phi}) = \mathbf{w}_k^T \boldsymbol{\phi}$ with $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$, and $\mathbf{w}_k = (w_{k0}, ..., w_{kM-1})^T$. Assume the following prior on the parameter vectors $\mathbf{w}_1, ..., \mathbf{w}_K$:

$$p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha) = \prod_{k=1}^K \frac{1}{(2\alpha)^M} e^{-|\mathbf{w}_k|_1/\alpha}.$$

Here, $|\mathbf{w}_k|_1$ denotes the L^1-norm of the vector $\mathbf{w}_k$, defined as $|\mathbf{w}_k|_1 = \sum_{m=0}^{M-1} |w_{km}|$.

Answer the following questions:

a) Write down the equation for the posterior $p(\mathbf{w}_1, ..., \mathbf{w}_K|\boldsymbol{\Phi}, \mathbf{T}, \alpha)$ in terms of the data likelihood $p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K)$ and the prior $p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha)$. You do not need to insert the actual distributions. **/1**

b) Compute the log-likelihood $\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K)$ and the log of the prior $\ln p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha)$. **/2**

c) Using your results in a) and b), show that the optimization problem corresponding to obtaining a Maximum A Posteriori (MAP) estimate for $\mathbf{w}_1, ..., \mathbf{w}_K$ is equivalent to performing L^1-regularized logistic regression for K classes, where we minimize the function

$$-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{\phi}_n) + \lambda \sum_{k=1}^{K}\sum_{m=0}^{M-1} |w_{km}|$$

with respect to $\mathbf{w}_1, ..., \mathbf{w}_K$, and with λ a regularization penalty parameter. How is λ related to α? **/3**

Solutions

a) The posterior distribution over the parameters is given by (1p)

$$p(\mathbf{w}_1, ..., \mathbf{w}_K | \boldsymbol{\Phi}, \mathbf{T}, \alpha) = \frac{p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K) p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha)}{p(\mathbf{T}|\boldsymbol{\Phi}, \alpha)}.$$

b) The log-likelihood is given by

$$\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K) = \ln \prod_{n=1}^{N}\prod_{k=1}^{K} y_k(\boldsymbol{\phi}_n)^{t_{nk}} = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{\phi}_n). \text{ (1 pt)}$$

The log of the prior is given by

$$\begin{aligned}\ln p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha) &= \ln \prod_{k=1}^{K} \frac{1}{(2\alpha)^M} e^{-|\mathbf{w}_k|_1/\alpha} = -KM \ln 2\alpha - \frac{1}{\alpha}\sum_{k=1}^{K} |\mathbf{w}_k|_1 \\ &= -KM \ln 2\alpha - \frac{1}{\alpha}\sum_{k=1}^{K}\sum_{m=0}^{M-1} |w_{km}| \text{ (1 pt)}\end{aligned}$$

c) The MAP estimate for $\mathbf{w}_1, ..., \mathbf{w}_K$ is obtained by maximizing the posterior distribution with respect to the parameters $\mathbf{w}_1, ..., \mathbf{w}_K$:

$$\begin{aligned}\mathbf{w}_1^{\text{MAP}}, ..., \mathbf{w}_K^{\text{MAP}} &= \arg\max_{\mathbf{w}_1,...,\mathbf{w}_K} p(\mathbf{w}_1, ..., \mathbf{w}_K|\boldsymbol{\Phi}, \mathbf{T}, \alpha) = \arg\min_{\mathbf{w}_1,...,\mathbf{w}_K} -\ln p(\mathbf{w}_1, ..., \mathbf{w}_K|\boldsymbol{\Phi}, \mathbf{T}, \alpha) \\ &= \arg\min_{\mathbf{w}_1,...,\mathbf{w}_K} -\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K) - \ln p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha), \text{ (2 pt)}\end{aligned}$$

where we have used the fact that $\frac{\partial}{\partial \mathbf{w}_j} -\ln p(\mathbf{T}|\boldsymbol{\Phi}, \alpha) = 0$ for $j = 1, ..., K$. Noting that the first term of the log-prior is independent of $\mathbf{w}_1, ..., \mathbf{w}_K$, we obtain

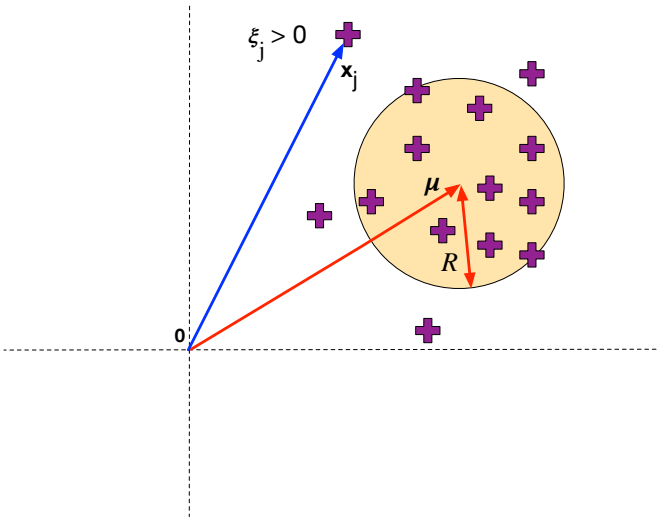
$$\mathbf{w}_1^{\text{MAP}}, ..., \mathbf{w}_K^{\text{MAP}} = \arg\min_{\mathbf{w}_1,...,\mathbf{w}_K} -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{\phi}_n) + \frac{1}{\alpha}\sum_{k=1}^{K}\sum_{m=0}^{M-1} |w_{km}|,$$

such that $\lambda = 1/\alpha$. (1 pt)

3 Outlier Detection

/14

We receive a dataset of N two-dimensional datapoints $\{\mathbf{x}_n\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^2$. See the figure below for an illustration of an example dataset with the purple crosses representing datapoints. We expect our data to lie on a disk with radius $R > 0$ and with its center at the position given by the vector $\boldsymbol{\mu}$. Due to noise some datapoints might fall outside of the disk and should be considered as outliers. Our goal is to find the radius R and the vector $\boldsymbol{\mu}$, such that the surface of the ring is minimized, and such that most datapoints lie on the surface of the disk. We introduce slack variables ξ_n for $n = 1, ..., N$, such that for all datapoints: $||\mathbf{x}_n - \boldsymbol{\mu}||^2 \leq R^2 + \xi_n$, with $\xi_n \geq 0$. A datapoint $\mathbf{x}_n$ that falls outside of the ring has a corresponding $\xi_n > 0$ (see $\mathbf{x}_j$ in figure below). In order to reduce the number of outliers we enforce a penalty for each datapoint with nonzero ξ_n.



To summarize, we want to minimize $\pi R^2 + C \sum_{n=1}^{N} \xi_n$, with hyperparameter $C > 0$ and the following constraints:

$$(\mathbf{1}) \quad ||\mathbf{x}_n - \boldsymbol{\mu}||^2 \leq R^2 + \xi_n \quad \text{for all} \quad n = 1, ..., N$$
$$(\mathbf{2}) \quad \xi_n \geq 0 \quad \text{for all} \quad n = 1, ..., N$$

a) Write down the primal Lagrangian function. Which variables are the primal variables? **/3**

b) Compute the derivatives of the Lagrangian with respect to the primal variables. Use these derivatives to derive conditions on the Lagrange multipliers. *Hint*: for the primal variable R it is easiest to take the derivative with respect to R^2. **/3**

c) Write down all of the KKT conditions. Do not consider the conditions computed at b) as KKT conditions. How many KKT conditions do we have in total? **/3**

d) Derive the dual Lagrangian of the problem and specify the optimization problem: with respect to which variables do you need to optimize, and do you need to maximize or minimize? Do not forget to list the conditions on the variables that you need to optimize with respect to! **/4**

e) Apply the kernel trick to the dual optimization problem. If you have not managed to write down a dual Lagrangian, you can also explain in words what applying the kernel trick means if you had a dual Lagrangian in the correct form. **/1**

Solutions

a)

$$\mathcal{L}(R, \boldsymbol{\mu}, \xi_n) = \pi R^2 + C \sum_{n=1}^{N} \xi_n + \sum_{n=1}^{N} \alpha_n (||\mathbf{x}_n - \boldsymbol{\mu}||^2 - R^2 - \xi_n) - \sum_{n=1}^{N} \beta_n \xi_n$$

(1p) for the right collection of terms.
(1p) for the right signs.
(1p) for naming the primal variables $R, \boldsymbol{\mu}, \xi_n$.

b)

$$\frac{\partial \mathcal{L}}{\partial R^2} = \pi - \sum_{n=1}^{N} \alpha_n = 0 \quad \rightarrow \quad \sum_{n=1}^{N} \alpha_n = \pi$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = -2 \sum_{n=1}^{N} \alpha_n (\mathbf{x}_n - \boldsymbol{\mu})^T = 0 \quad \rightarrow \quad \boldsymbol{\mu} = \frac{\sum_{n=1}^{N} \alpha_n \mathbf{x}_n}{\sum_{n=1}^{N} \alpha_n} = \frac{1}{\pi} \sum_{n=1}^{N} \alpha_n \mathbf{x}_n$$

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \beta_n = 0 \quad \rightarrow \quad \alpha_n = C - \beta_n.$$

(0.5p) for each derivative. (1p) for setting derivatives equal to zero.

c)

$$\begin{aligned} \alpha_n &\geq 0 \\ ||\mathbf{x}_n - \boldsymbol{\mu}||^2 - R^2 - \xi_n &\leq 0 \\ \alpha_n (||\mathbf{x}_n - \boldsymbol{\mu}||^2 - R^2 - \xi_n) &= 0 \end{aligned}$$

$$\begin{aligned} \beta_n &\geq 0 \\ \xi_n &\geq 0 \\ \beta_n \xi_n &= 0 \end{aligned}$$

for all $n = 1, ..., N$. (2p) for all constraints with correct signs, and (1p) for the number $2 \cdot 3N = 6N$.

d) Collecting all terms that correspond to each primal variable and using the conditions derived in (b) leads to

$$\begin{aligned}\tilde{\mathcal{L}} =& R^2(\pi - \sum_n \alpha_n) + \sum_n \xi_n(C - \alpha_n - \beta_n) + \sum_n \alpha_n(\mathbf{x}_n - \boldsymbol{\mu})^T(\mathbf{x}_n - \boldsymbol{\mu}) \\ =& \sum_n \alpha_n(\mathbf{x}_n - \frac{1}{\pi}\sum_m \alpha_m \mathbf{x}_m)^T(\mathbf{x}_n - \frac{1}{\pi}\sum_l \alpha_l \mathbf{x}_l) \\ =& \sum_n \alpha_n \mathbf{x}_n^T\mathbf{x}_n - \frac{2}{\pi}\sum_n\sum_m \alpha_n\alpha_m \mathbf{x}_n^T\mathbf{x}_m + \frac{1}{\pi^2}\sum_n \alpha_n \sum_m\sum_l \alpha_m\alpha_l \mathbf{x}_m^T\mathbf{x}_l \\ =& \sum_n \alpha_n \mathbf{x}_n^T\mathbf{x}_n - \frac{1}{\pi}\sum_n\sum_m \alpha_n\alpha_m \mathbf{x}_n^T\mathbf{x}_m. \text{(2p)}\end{aligned}$$

In going from the second to last to the last line we have used $\sum_n \alpha_n = \pi$. The dual optimization problem then consists of maximizing $\tilde{\mathcal{L}}$ with respect to the dual variables $\{\alpha_n\}_{n=1}^N$. (1p)

The resulting conditions for α_n can be constructed by combining the conditions in b) and c). From the KKT conditions in c) we know that $\alpha_n \geq 0$. In b) we also found that $\alpha_n = C - \beta_n$, and in c) we saw that $\beta_n \geq 0$, hence $\alpha_n \leq C$. Together with the first condition that we found in c) this leaves us with the following (1p).

$$\sum_n \alpha_n = \pi$$
$$0 \leq \alpha_n \leq C.$$

e) Replace all inner products $\mathbf{x}_n^T\mathbf{x}_m$ with a valid kernel function $k(\mathbf{x}_n, \mathbf{x}_m)$. The resulting dual Lagrangian then yields

$$\tilde{\mathcal{L}} = \sum_n \alpha_n k(\mathbf{x}_n, \mathbf{x}_n) - \frac{1}{\pi}\sum_n\sum_m \alpha_n\alpha_m k(\mathbf{x}_n, \mathbf{x}_m). \text{ (1p)}$$

4 Mixture of Exponentials

/7

Our task is to cluster series of measurements containing the time differences between events at which radioactive particles decay. We are given an unlabelled dataset $X = \{\boldsymbol{x}_n\}_{n=1}^N$, where each $\boldsymbol{x}_n$ represents a series of time differences measured. Each $\boldsymbol{x}_n$ is a vector of size D containing real numbers such that $x_{ni} \geq 0$ for $i = 1, .., D$. D is the total number of time differences measured for each measurement. We assume that there are K different sources of decaying particles, and that the measurements are generated as follows:

i) The sources are represented by a discrete latent variable $z \in \{1, ..., K\}$ with probability distribution $p(z) = \prod_{k=1}^K \pi_k^{I[z=k]}$ and $\sum_{k=1}^K \pi_k = 1$. Here, $I[z = k]$ is the indicator function. The parameters $\pi_k \geq 0$ represent the prior probabilities for each source k being present, and are unknown, so they need to be learned.

ii) For a real valued vector $\boldsymbol{x}$ with elements larger or equal to zero, that corresponds to a measurement of source $z = k$, each x_i $(i = 1, ..., D)$ is sampled independently from an Exponential distribution with parameter λ_k:

$$p(x_i|z = k) = \lambda_k e^{-\lambda_k x_i} \ .$$

The parameters $\lambda_k > 0$ are so-called rate parameters, and need to be learned.

Answer the following questions.

a) How many parameters does our model contain? Indicate how this number depends on K, D, N. **/1**

b) Compute the probability of a measurement $\boldsymbol{x}_n$ conditioned on source k: $p(\boldsymbol{x}_n|z = k)$. Compute the marginal probability of $\boldsymbol{x}_n$ under this model: $p(\boldsymbol{x}_n)$. Your answers should be functions of the model parameters and the datapoints. **/2**

c) Compute the responsibility (or posterior) $r_{nk} = p(z = k|\boldsymbol{x}_n)$ of a measurement with feature vector $\boldsymbol{x}_n$ containing time differences between decays of source k. **/1**

d) For deriving an EM algorithm for a Mixture of Exponentials, it is convenient to express the log-likelihood in a different way. This is called the *expected complete log-likelihood*:

$$\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X})}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\mu})\right] = \ln\left(\prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{r_{nk}} p(\boldsymbol{x}_n|z_n = k)^{r_{nk}}\right)$$
$$= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\ln \pi_k + D\ln\lambda_k - \lambda_k\sum_{i=1}^{D} x_{ni}\right) . \qquad (1)$$

From the expression in Eq. (1), obtain an update rule for each parameter λ_k as a function of the responsibilities r_{nk}. Explain in words how you would use this update rule in the EM algorithm. **/3**

Solutions

a) There are two sets of parameters, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$. Both $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ are vectors of dimension K (one per source). Therefore the total number is $2K$ (1p). There is no dependence on N and D. To be more precise, since $\sum_{k=1}^{K}\pi_k = 1$, we may notice that there is no need to store all K μ_k parameters, but it suffices to have $K - 1$; the total is then $2K - 1$.

b) Because of independence we write:

$$p(\boldsymbol{x}_n|z = k) = \prod_{i=1}^{D}\lambda_k e^{-\lambda_k x_{ni}} \text{ .(1pt)}$$

The marginal:

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K}\pi_k\prod_{i=1}^{D}\lambda_k e^{-\lambda_k x_{ni}} \text{ .(1pt)}$$

c) By Bayes theorem:

$$r_{nk} = p(z = k|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|z = k)p(z = k)}{p(\boldsymbol{x}_n)}$$
$$= \frac{\pi_k\prod_{i=1}^{D}\lambda_k e^{-\lambda_k x_{ni}}}{\sum_{j=1}^{K}\pi_j\prod_{i=1}^{D}\lambda_j e^{-\lambda_j x_{ni}}} .$$

(1pt) for Bayes rule.

d) It is not necessary to show *all* of the steps below; these solutions show them all for the sake of clarity. We can differentiate with respect to λ_k.

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_k} \sum_{n=1}^{N} \sum_{k'=1}^{K} r_{nk'} \left(\ln \pi_{k'} + D \ln \lambda_{k'} - \lambda_{k'} \sum_{i=1}^{D} x_{ni} \right) \\
&= \sum_{n=1}^{N} r_{nk} \frac{\partial}{\partial \lambda_k} \left(D \ln \lambda_k - \lambda_k \sum_{i=1}^{D} x_{ni} \right) \\
&= \sum_{n=1}^{N} r_{nk} \left(\frac{D}{\lambda_k} - \sum_{i=1}^{D} x_{ni} \right) = 0 \ . \text{(1pt)}
\end{aligned}$$

Reshuffling, we obtain

$$\lambda_k = \frac{\sum_{n=1}^{N} r_{nk}}{\frac{1}{D} \sum_{i=1}^{D} \sum_{n=1}^{N} r_{nk} x_{ni}} \ , \text{(1pt)}$$

Note that this makes sense since λ_k is the rate parameter of the exponential distribution, and as such λ_k should have units of inverse time. The above result shows that $1/\lambda_k$ is equal to the weighted mean of the time differences in between decays in each datapoint, with weighting coefficients given by the responsibilities that source k takes for the datapoints.

Updating all of the λ_k is part of the Maximization step in the Expectation-Maximization algorithm, where we update all the parameters and we keep the posterior fixed; the Maximization step would also include an update for $\boldsymbol{\pi}$. In the Expectation-step, we re-compute the posterior probabilities r_{nk}, while keeping all the parameters fixed .(1pt).

5 Principle Component Analysis

/7

Suppose we have a dataset consisting of N datapoints: $\{\mathbf{x}_n\}_{n=1}^{N}$ of D-dimensional vectors. The data is collected in a matrix $\mathbf{X}$ of size $N \times D$, such that the n-th row is given by $\mathbf{x}_n^T$. Consider the following steps of the procedure for PCA for dimensionality reduction:

step 1 Center $\mathbf{X}$, with a centered data matrix $\hat{\mathbf{X}}$ as a result, with its n-th row given by $\hat{\mathbf{x}}_n^T$. Here, $\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}$ with $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$.

step 2 Compute the sample covariance matrix $\mathbf{S}$ of the centered dataset.

step 3 Solve the eigenvalue problem $\mathbf{S} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$, where $\mathbf{U}$ is a D by D matrix with the eigenvectors of $\mathbf{S}$ as its columns, such that $\mathbf{u}_k$ is the eigenvector corresponding to the k-th largest eigenvalue. The eigenvectors are orthonormal: $\mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$, with $\delta_{ij} = 1$ if $i = j$ and zero otherwise. The matrix $\boldsymbol{\Lambda}$ is only nonzero on the diagonal, with the eigenvalues λ_k of $\mathbf{S}$ on its diagonal, such that λ_k is the k-th largest eigenvalue.

step 4 Choose the $M < D$ eigenvectors corresponding to the M largest eigenvalues and use them to construct the matrix $\mathbf{U}_M = [\mathbf{u}_1, ..., \mathbf{u}_M]$.

step 5 Project the datapoints onto an M-dimensional subspace using $\mathbf{z}_n = \mathbf{U}_M^T (\mathbf{x}_n - \bar{\mathbf{x}}) = \mathbf{U}_M^T \hat{\mathbf{x}}_n$. If we collect the projected dataset in a matrix $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_N]^T$, then the projection is given by $\mathbf{Z} = \hat{\mathbf{X}} \mathbf{U}_M$.

Answer the following questions:

a) Write down the sample covariance matrix $\mathbf{S}$ of the centered dataset in terms of $\hat{\mathbf{X}}$. What is the shape of $\mathbf{S}$? . **/2**

b) If you are interested in retaining 65% of the variance of the original data with a projection onto an $M < D$ dimensional subspace using PCA, how do you choose M? Write an explicit equation that you can use to determine M. **/2**

c) If we are not interested in dimensionality reduction, but only in making sure the projected dataset has unit covariance, we need to choose $M = D$ and adjust step 5. What is the name of this procedure? Write down the explicit equation for the operation, and show that the covariance matrix of the projected dataset is the identity matrix. **/3**

Solutions

a) $\mathbf{S} = \frac{1}{N}\hat{\mathbf{X}}^T\hat{\mathbf{X}}$ (1p) and is a D by D matrix (1p).

b) The variance of the original data is given by $\sum_{k=1}^{D}\lambda_k$, and the variance of the projected data is given by $\sum_{k=1}^{M}\lambda_k$. To choose the optimal M for retaining 65% of the original variance in the projected dataset set we thus need to choose the smallest M such that

$$\frac{\sum_{k=1}^{M}\lambda_k}{\sum_{k=1}^{D}\lambda_k} \geq 0.65 \text{ . (1p)}$$

c) This is called whitening/sphering (1p). In order to whiten the data we need to alter step 5 such that $\mathbf{z}_n = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{x}_n - \bar{\mathbf{x}})_n$ or $\mathbf{Z} = \hat{\mathbf{X}}\mathbf{U}\mathbf{\Lambda}^{-1/2}$ (1p). The resulting covariance matrix is given by

$$\begin{aligned}\frac{1}{N}\mathbf{Z}^T\mathbf{Z} =& \frac{1}{N}\mathbf{\Lambda}^{-1/2}\mathbf{U}^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{S}\mathbf{U}\mathbf{\Lambda}^{-1/2} \\ =& \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} = \mathbb{1} \text{ . (1p)}\end{aligned}$$

We have made use of the orthonormality of the eigenvectors in $\mathbf{U}$ such that $\mathbf{U}^T\mathbf{U} = \mathbb{1}$.