UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Final Exam
Date: October 23, 2017
Time: 13:00-16:00

Number of pages: 11 (including front page)
Number of questions: 4
Maximum number of points to earn: 47
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

/17

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

   - [X] For regression with targets $t \in \mathbb{R}$, a suitable activation function for the output unit is $f(x) = x$.
   - [ ] For classification with $K > 2$ mutually exclusive classes we need K output units with activation functions $f(x) = \frac{1}{1+e^{-x}}$.
   - [X] For classification with binary targets we can use 1 output unit with activation function $f(x) = \frac{1}{1+e^{-x}}$.
   - [ ] For regression with targets $t \geq 0$, a suitable activation function for the output unit is $f(x) = \tanh(x)$.

2. Consider a neural network with L layers, M hidden units for each hidden layer, and K output units. Which of the following statements are correct? /1

   - [X] Increasing L, while keeping M fixed, increases the risk of overfitting.
   - [X] Adding a weight decay term to the loss function for all of the weights in the neural network will help reduce overfitting.
   - [ ] You can decrease the risk of overfitting by decreasing K.
   - [ ] The activation functions for the hidden units need to be the same as the activation function for the output units.

3. Which of the following statements about training a neural network with stochastic gradient descent (SGD) are correct? Assume the SGD is performed by sampling single data points, not with minibatches. We denote the error function/loss function as $E = \sum_{n=1}^{N} E_n$ for N datapoints. /1

   - [X] In the forward propagation step we take a single data point as input, and then compute all of the activations of the hidden and output units.
   - [X] In the backpropagation step the chain rule is used to compute the derivatives $\frac{\partial E_n}{\partial w_{ij}^{(l)}}$, for all weights $w_{ij}^{(l)}$. Here, l indicates the layer of the corresponding weight.
   - [ ] In backpropagation we first compute the gradient of the error function with respect to the weights in the output layer L, and then update those weights. We then use these updated weights to compute the derivatives and updates of the weights in the last hidden layer $L - 1$.
   - [ ] Stochastic gradient descent is more sensitive to get stuck in local minima than full batch gradient descent.

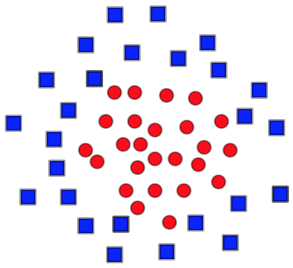
4. Which of the following statements about K-means clustering and Gaussian Mixture Models (GMM) are correct? **/1**

   - [X] In the K-means algorithm hard cluster assignments are made, whereas in GMM soft cluster assignments are made.
   - [ ] K-means is insensitive to feature scaling.
   - [X] K-means tries to model clusters with spherical shapes.
   - [ ] For Gaussian Mixture Models you do not need to choose the number of clusters K in advance. For K-means you do need to choose K.

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

   - [X] The specific form of the updates for the means, covariances and the cluster prior probabilities/mixture components can be derived by computing the maximum likelihood estimates for these three parameters respectively.
   - [X] In the EM algorithm, the E step corresponds to computing the responsibilities with all other parameters fixed.
   - [ ] The M step consists of updating the means and the covariances, but not the cluster mixture components.
   - [ ] The EM algorithm always converges to the global minimum.

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



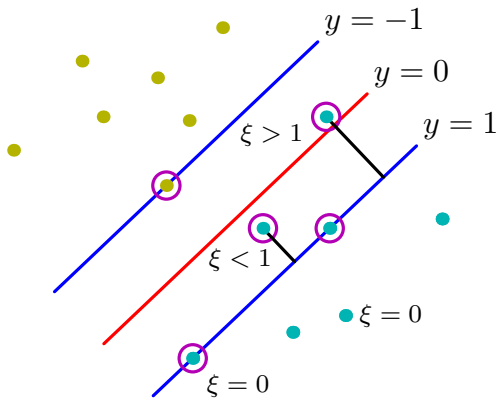
/1

   - [X] Neural networks.
   - [ ] Logistic regression with linear features.
   - [ ] Linear Discriminant Analysis with linear features
   - [X] Support Vector Machines (with any kernel).

7. 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 7 yellow circles correspond to data points with labels $t_n = -1$, and the 7 green data represent the data points with labels $t_n = +1$.



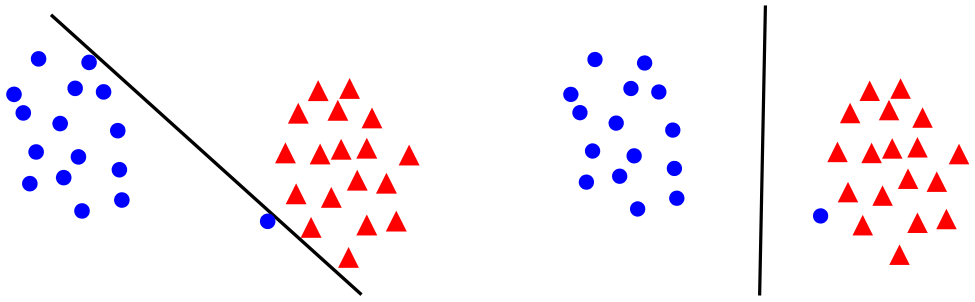
Which of the following statements are correct? **/1**

- [X] For $C \to \infty$ the number of support vectors is equal to 2. (should have been " at least 2")
- [ ] For $C \to \infty$ the number of support vectors is equal to 1.
- [ ] For $C = 0$ the number of support vectors is equal to 2.
- [X] For $C = 0$ all datapoints will become support vectors.

8. Consider a Maximum Margin classifier with soft margins for a dataset that is not linearly separable. We aim to minimize $C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$ with respect to $\mathbf{w}, \xi_n$ and b, subject to the constraints $t_n(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) - b) \geq 1 - \xi_n$. Assume the vector $\boldsymbol{\phi}(\mathbf{x}_n)$ is a nonlinear mapping of vector $\mathbf{x}_n$. Which of the following statements are correct? **/1**

- [X] For large C we expect a more complex decision boundary than for small C.
- [ ] For large C we expect a less complex decision boundary than for small C.
- [X] The classifier is more sensitive to outliers for large C than for small C.
- [ ] The classifier is not sensitive to outliers.

9. Consider the following figures depicting a dataset with datapoints from two classes corresponding to the blue circles and the red triangles:



(a) (b)

The black lines correspond to decision boundaries constructed using (different) Maximum Margin classifiers. Which of the following statements about the above figures are correct: **/1**

- [X] It is likely that the decision boundary in (a) is determined by a Maximum Margin classifier with a hard margin.
- [ ] The decision boundary in (a) is more likely to lead to a good generalization performance than the decision boundary shown in (b).
- [X] It is possible that the decision boundary in (a) is determined by a Maximum Margin classifier with a soft margin and an extremely high penalty for misclassifications.
- [X] It is likely that the decision boundary in (b) is determined by a Maximum Margin classifier with a soft margin and a low penalty for misclassifications.

10. The Lagrangian function for the Maximum Margin classification problem for a linearly separable dataset $\{\mathbf{x}_n, t_n\}_{n=1}^N$ is given by

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{n=1}^{N} a_n\{t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1\},$$

where $\mathbf{a} = (a_1, ..., a_N)^T$ is a vector of Lagrange multipliers for the constraints $t_n(\mathbf{w}^T\mathbf{x}_n + b) - 1 \geq 0$ with $n = 1, ..., N$ and $a_n \geq 0$. The dual representation of the Maximum Margin problem is given by maximizing the following equation with respect to a_n, for $n = 1, ..., N$:

$$\widetilde{L} = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \mathbf{x}_n^T \mathbf{x}_m. \quad (1)$$

With the constraints $a_n \geq 0$ and $\sum_{n=1}^{N} a_n t_n = 0$. Which of the following statements are correct? **/1**

- [X] The dual formation allows the use of the kernel trick.
- [X] The kernel trick consists of replacing all $\mathbf{x}_n^T\mathbf{x}_m$ with kernels $k(\mathbf{x}_n, \mathbf{x}_m)$ in Eq. (1).
- [ ] The support vectors are those datapoints for which $a_n = 0$.
- [X] The support vectors are those datapoints for which $a_n > 0$.

11. Which of the following statements about Gaussian processes are correct? **/1**

- [X] The Gaussian Process model is a nonparametric model.
- [X] Consider a Gaussian Process model for regression that has been trained on N datapoints $\{\mathbf{x}_n, t_n\}_{n=1}^N$. The predictive distribution for a new target t_{N+1} with input vector $\mathbf{x}_{N+1}$, is a Gaussian whose mean and variance both depend on $\mathbf{x}_{N+1}$.
- [ ] Gaussian processes can not be used for classification.
- [ ] There are no hyperparameters in Gaussian processes.

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

- [X] Given two independent, normally distributed random variables X and Y, the random variable $Z = X + Y$ is also normally distributed.
- [ ] In Gaussian Processes for regression we assume the targets t_n are generated by $t_n = y_n \times \varepsilon_n$ where ε_n is sampled independently for each datapoint from $\mathcal{N}(\varepsilon|0, \beta^{-1})$.
- [X] In Gaussian Processes for regression we assume the targets t_n are generated by $t_n = y_n + \varepsilon_n$ where ε_n is sampled independently for each datapoint from $\mathcal{N}(\varepsilon|0, \beta^{-1})$.
- [X] Consider two random variables $\mathbf{t} \in \mathrm{R}^N$ and $\mathbf{y} \in \mathrm{R}^N$, such that $p(\mathbf{t}|\mathbf{y}) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbb{1}_N)$ and $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$. Here $\mathbb{1}_N$ is the identity matrix of size $N \times N$, and K is a $N \times N$ positive definite covariance matrix. Then $p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\mathrm{d}\mathbf{y}$ is also a multivariate Gaussian distribution.

13. Which of the following problems or algorithms are used for unsupervised learning: **/1**

- [X] Outlier detection with support vector machines.
- [X] Principle Component Analysis (PCA).
- [X] Gaussian Mixture Models.
- [ ] Linear Discriminant Analysis.

14. Indicate which of the following statements about Principle Component Analysis (PCA) are correct: **/1**

- [ ] PCA tries to find a nonlinear projection of the data such that the variance in the projected space is maximal.
- [X] The principle components correspond to eigenvectors of the covariance matrix of the data.
- [ ] Assume that the data is D-dimensional. In order to project the data down to M dimensions ($M < D$), we need to find the M eigenvectors of the data covariance matrix with the M smallest eigenvalues.
- [X] After projecting D-dimensional data on an M dimensional subspace with PCA ($M < D$), the variance of the projected data is given by $\sum_{j=1}^{M} \lambda_j$. Here, λ_j with $j = 1, ..., M$ are the M largest eigenvalues of the data covariance matrix.

15. The covariance matrix $\mathbf{S}$ of a data set $\mathcal{D} = \{\mathbf{X}\}$ with $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]^T$ (i.e. an N by D matrix, where D is the number of features) is given by: **/1**

- [X] $\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})^T$.
- [ ] $\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})^T(\mathbf{x_n} - \frac{1}{N}\sum_{n=1}^{N}\mathbf{x_n})$.
- [ ] $\frac{1}{N}\mathbf{X}\mathbf{X}^T$, if the data is zero-centered (i.e. mean subtracted).
- [X] $\frac{1}{N}\mathbf{X}^T\mathbf{X}$, if the data is zero-centered (i.e. mean subtracted).

16. Which of the following statements about Principle Component Analysis (PCA) are correct? **/1**

- [X] PCA can be used for 2D visualization of high dimensional data.
- [X] After applying PCA to a dataset, the covariance matrix of the projected data is diagonal.
- [ ] You aim at projecting D-dimensional data on an M dimensional subspace with PCA ($M < D$), such that at least 80% of the variance of the data is preserved. The correct way of choosing M is such that $\left(\sum_{j=1}^{M}\lambda_j\right) / \left(\sum_{i=1}^{D}\lambda_i\right) < 0.8$.
- [ ] None of the above.

17. Which of the following statements about Random Forest classifiers are correct? **/1**

- [ ] Random Forest classifiers use boosting.
- [X] Random Forest classifiers use bootstrapping.
- [X] Random Forest classifiers average results from multiple models.
- [X] Random Forest classifiers use bagging.

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 47, meaning that the final grade is computed as $10 \times \frac{\#\text{points}}{47}$. The second option in multiple choice question 3 is not taken into account as the partial derivative sign in the denominator was missing in the exam. The first option in multiple choice question 7 is also not taken into account, as in general the number of support vectors is at least 2 for the $C \to \infty$ limit, and in this case the number of support vectors is larger than 2. However you can only see this by drawing carefully.

General remarks

The exercises below have subquestions that are mostly independent. If you get stuck at one subquestion, don't stop but try to solve the next ones!

2 K-means clustering

/7

Consider the K-Means algorithm with the following cost function:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2$$

with data points $\mathbf{x}_n$, assignment indicator variables $r_{nk} \in \{0, 1\}$ and cluster centers $\boldsymbol{\mu}_k$.

a) K-means can be understood as a sequence of EM iterations. Give the equations that describe the updates for the E-step and the M-step respectively. When has the algorithm converged? **/3**

b) Suppose we have a K-means clustering algorithm with three cluster centroids given by $\boldsymbol{\mu}_1 = (1, 2)^T$, $\boldsymbol{\mu}_2 = (-3, 0)^T$, $\boldsymbol{\mu}_3 = (4, 2)^T$. Consider the following data point $\mathbf{x}_i = (-1, 2)^T$ in the training set. The next update is the E-step where data points are assigned to clusters. To which cluster centre will the data point $\mathbf{x}_i$ be assigned? **/2**

c) Consider the case where you have an image consisting of 1×10^6 pixels, with each pixel represented by a vector of size 3 with the RGB values. You want to perform image compression by applying a K-means clustering algorithm to the image. What does K represent in this setting? And what do the cluster centroids represent after the K-means clustering has converged? **/2**

Solutions

a) The E-step:

$$r_{nk} = \begin{cases} 1 & \text{if} \quad k = \arg\min_j ||\mathbf{x}_n - \boldsymbol{\mu}_j|| \\ 0 & \text{otherwise} \end{cases} \quad \text{(1p)}$$

The M-step:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}. \quad \text{(1p)}$$

The algorithm as converged when the cluster assignments no longer change (1p).

b) The three cluster centroids are given by $\boldsymbol{\mu}_1 = (1, 2)^T$, $\boldsymbol{\mu}_2 = (-3, 0)^T$, $\boldsymbol{\mu}_3 = (4, 2)^T$. The data point $\mathbf{x}^{(i)} = (-1, 2)^T$ will be assigned to the closest centroid in the next E-step. The squared distance to each of the clusters is given by

$$||\mathbf{x}^{(i)} - \boldsymbol{\mu}_1||^2 = \sum_{j=1}^{2}(x_j^{(i)} - \mu_{1,j})^2 = (-2)^2 + 0^2 = 4$$
$$||\mathbf{x}^{(i)} - \boldsymbol{\mu}_2||^2 = (2)^2 + 2^2 = 8$$
$$||\mathbf{x}^{(i)} - \boldsymbol{\mu}_3||^2 = (-5)^2 + 0 = 25.$$

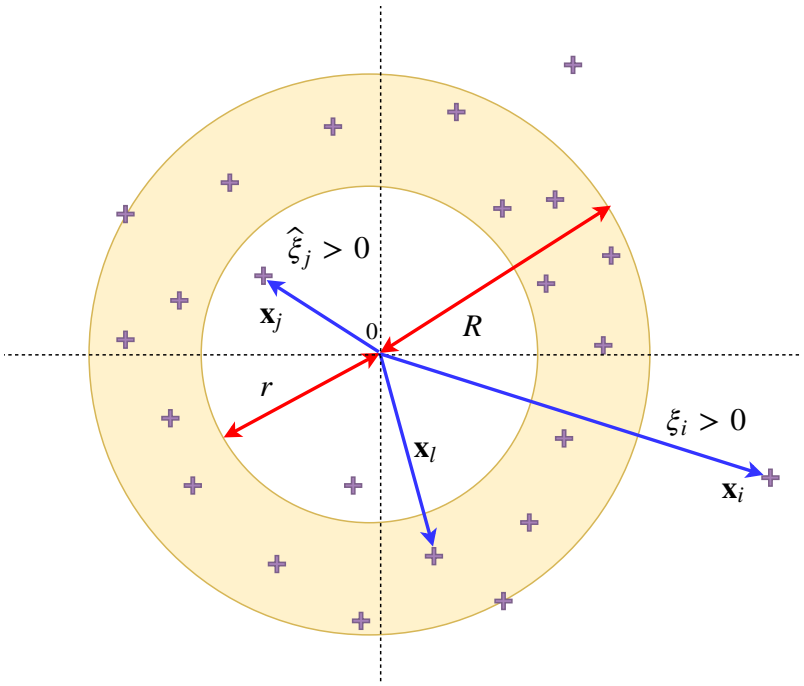
Cluster centroid $\boldsymbol{\mu}_1$ is thus the closest to the datapoint, and the datapoint $\mathbf{x}^{(i)}$ will be assigned to this cluster in the next E-step. (2)

c) K represents the number of colors that remain in the compressed image (1p). After convergence of the K-means clustering, the cluster centroids represent the RGB values of the resulting K colors that are used in the compressed image (1p).

3 Outlier detection

/11

We receive a dataset of N two-dimensional datapoints $\{\mathbf{x}_n\}_{n=1}^N$ with $\mathbf{x}_n \in \mathbb{R}^2$. The dataset has been centered around the origin $(0, 0)$. 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 ring with inner radius r and outer radius R $(R > r)$. Due to noise some datapoints might fall outside of the ring and should be considered as outliers. Our goal is thus to find the radii r and R, such that the surface of the ring is minimized, and such that most datapoints lie on the surface of the ring. We introduce slack variables ξ_n and $\widehat{\xi}_n$ for $n = 1, ..., N$, such that for all datapoints: $||\mathbf{x}_n||^2 \leq R^2 + \xi_n$, and $||\mathbf{x}_n||^2 \geq r^2 - \widehat{\xi}_n$, with $\xi_n \geq 0$ and $\widehat{\xi}_n \geq 0$. A datapoint $\mathbf{x}_n$ that falls outside of the ring either has $\xi_n > 0$ (see $\mathbf{x}_i$ in figure below), or $\widehat{\xi}_n > 0$ (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 or $\widehat{\xi}_n$.



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

(**1**) $||\mathbf{x}_n||^2 \leq R^2 + \xi_n$ for all $n = 1, ..., N$
(**2**) $||\mathbf{x}_n||^2 \geq r^2 - \widehat{\xi}_n$ for all $n = 1, ..., N$
(**3**) $\xi_n \geq 0$ for all $n = 1, ..., N$
(**4**) $\widehat{\xi}_n \geq 0$ for all $n = 1, ..., N$
(**5**) $\widehat{\xi}_n \leq r^2$ for all $n = 1, ..., N$
(**6**) $R^2 \geq r^2$

a) Write down the Lagrangian function. Use $\{\alpha_n\}_{n=1}^N$ and $\{\widehat{\alpha}_n\}_{n=1}^N$ as Lagrange multipliers for conditions (**1**) and (**2**) respectively, $\{\beta_n\}_{n=1}^N$ for condition (**3**), $\{\widehat{\beta}_n\}_{n=1}^N$ for condition (**4**), $\{\gamma_n\}_{n=1}^N$ for condition (**5**), and δ for condition (**6**). Which variables are the primal variables? **/3**

b) Compute the derivatives of the Lagrangian with respect to the four primal variables. Use these derivatives to derive four conditions on the Lagrange multipliers. *Hint*: for the primal variables r and R it is easiest to take derivatives with respect to r^2 and 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? *Hint*: each separate constraint listed in the introduction corresponds to three KKT conditions. **/3**

d) Explain in words how you would derive the dual representation of the problem. You do not have to perform the actual derivation. **/2**

Solutions

a)

$$\mathcal{L}(R, r, \xi_n, \widehat{\xi}_n) = \pi(R^2 - r^2) + C\sum_{n=1}^{N}(\xi_n + \widehat{\xi}_n) + \sum_{n=1}^{N}\alpha_n(||\mathbf{x}_n||^2 - R^2 - \xi_n)$$
$$- \sum_{n=1}^{N}\widehat{\alpha}_n(||\mathbf{x}_n||^2 - r^2 + \widehat{\xi}_n) - \sum_{n=1}^{N}\beta_n\xi_n - \sum_{n=1}^{N}\widehat{\beta}_n\widehat{\xi}_n + \sum_{n=1}^{N}\gamma_n(\widehat{\xi}_n - r^2) - \delta(R^2 - r^2)$$

(1p) for the right collection of terms
(1p) for the right signs
(1p) for naming the primal variables $R, r, \xi_n, \widehat{\xi}_n$

b)

$$\frac{\partial\mathcal{L}}{\partial R^2} = \pi - \sum_{n=1}^{N}\alpha_n - \delta = 0$$
$$\frac{\partial\mathcal{L}}{\partial r^2} = -\pi + \sum_{n=1}^{N}\widehat{\alpha}_n - \sum_{n=1}^{N}\gamma_n + \delta = 0$$
$$\frac{\partial\mathcal{L}}{\partial\xi_n} = C - \alpha_n - \beta_n = 0$$
$$\frac{\partial\mathcal{L}}{\partial\widehat{\xi}_n} = C - \widehat{\alpha}_n - \widehat{\beta}_n + \gamma_n = 0$$

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

c)

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

$$\begin{aligned}
\widehat{\alpha}_n &\geq 0 \\
||\mathbf{x}_n||^2 - r^2 + \widehat{\xi}_n &\geq 0 \\
\widehat{\alpha}_n(||\mathbf{x}_n||^2 - r^2 + \widehat{\xi}_n) &= 0
\end{aligned}$$

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

$$\begin{aligned}
\widehat{\beta}_n &\geq 0 \\
\widehat{\xi}_n &\geq 0 \\
\widehat{\beta}_n \widehat{\xi}_n &= 0
\end{aligned}$$

$$\begin{aligned}
\gamma_n &\geq 0 \\
\widehat{\xi}_n - r^2 &\leq 0 \\
\gamma_n(\widehat{\xi}_n - r^2) &= 0
\end{aligned}$$

$$\begin{aligned}
\delta &\geq 0 \\
R^2 - r^2 &\geq 0 \\
\delta(R^2 - r^2) &= 0
\end{aligned}$$

(2p) for all constraints, with and (1p) for the number $3 + 5 \cdot 3N$.

d) The dual formulation of the problem can be obtained by eliminating the primal variables from the Lagrangian through the use of the four conditions obtained in (b). Collect all terms in the Lagrangian with respect to each primal variable, then you will see the conditions in (b) appearing, so that the primal variables will dropout. The rest of the terms together make up the dual formulation. (2p).

4 Mixture of Bernoullis and dimensionality reduction

/12

Our task is to cluster together Facebook pages with respect to their topics by looking at their description (text). We are given an unlabelled dataset $X = \{\boldsymbol{x}_n\}_{n=1}^N$, where each $\boldsymbol{x}_n$ represents a page. Each $\boldsymbol{x}_n$ is a binary vector of size D, with D the size of the dictionary of all possible words that we consider. For each $i = 1, \ldots, D$, $x_{ni} = 1$ if the i-th word of the dictionary appears (at least once) in the description of page n, and otherwise $x_{ni} = 0$. We assume that there are K topics, and that the descriptions are generated as follows:

i) The topics 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 cluster k, and are unknown, so they need to be learned.

ii) For a binary feature vector $\mathbf{x}$ that corresponds to a page with topic $z = k$, each x_i ($i = 1, ..., D$) is sampled independently from a Bernoulli distribution with parameter μ_{ki}:

$$p(x_i|z = k) = \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i} \quad .$$

The parameters μ_{ki} 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 single page $\boldsymbol{x}_n$ conditioned on topic 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) $r_{nk} = p(z = k|\mathbf{x}_n)$ of a topic k generating a page with feature vector $\boldsymbol{x}_n$. **/1**

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

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})} \left[\ln p(X, Z|\boldsymbol{\pi}, \boldsymbol{\mu})\right] &= \ln \left(\prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{r_{nk}} p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)^{r_{nk}} \right) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left(\ln \pi_k + \sum_{i=1}^{D} x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki}) \right) \quad . \end{aligned} \tag{2}$$

From the expression in Eq. (2), obtain an update rule for each parameter μ_{ki} as a function of the responsibilities r_{nk}. You may assume that $0 < \mu_{ki} < 1$ for all k and i. *Hint*: the result is of the same form as for GMM's:

$$\mu_{ki} = \frac{\sum_{n=1}^{N} r_{nk} x_{ni}}{\sum_{n=1}^{N} r_{nk}}, \quad \text{for all} \quad k, i$$

Explain in words how you would use this update rule in the EM algorithm. **/3**

e) Assume now that $N = 100 \cdot 10^6$. Motivate why the update rule for μ_{ki} found before may not be appropriate and *write explicitly* a better type of update. *Hint*: think of stochastic gradient descent updates. **/2**

f) Assume now that $D = 50 \cdot 10^3$, which we consider a too high dimensional problem for our computing resources. Name a preprocessing technique that we could use for transforming the data into a more maneagable representation, while still retaining most of the information. Write explicitly all the steps of the chosen algorithm. Can we still model our problem with a Mixture of Bernoulli after applying the transformation? **/3**

Solutions

a) There are two sets of parameters, $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$. $\boldsymbol{\pi}$ is a vector of dimension K (one per topic), while there is a μ_{ki} for each topic and word. Therefore the total number is $K + KD$ (1p). There is no dependence on N since this is a parametric model. To be more precise, since $\sum_{k=1}^{K} \pi_k = 1$, we may notice that there is no need to store all K parameters, but it suffices to have $K - 1$; the total is then $K + KD - 1$.

b) Because of independence we write:

$$p(\boldsymbol{x}_n|z = k) = \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}\right)^{1-x_{ni}} \quad .(1\text{pt})$$

The marginal:

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}\right)^{1-x_{ni}} \quad .(1\text{pt})$$

c) By Bayes theorem:

$$\begin{aligned} r_{nk} = p(z = k|\boldsymbol{x}_n) &= \frac{p(\boldsymbol{x}_n|z = k)p(z = k)}{p(\boldsymbol{x}_n)} \\ &= \frac{\pi_k \prod_{i=1}^{D} \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}\right)^{1-x_{ni}}}{\sum_{j=1}^{K} \pi_j \prod_{i=1}^{D} \mu_{ji}^{x_{ni}} \left(1 - \mu_{ji}\right)^{1-x_{ni}}} \quad . \end{aligned}$$

(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 μ_{ki}.

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

Assuming that neither μ_{ki} nor $1 - \mu_{ki}$ are equal to 0, we can simplify to:

$$\sum_{n=1}^{N} r_{nk} x_{ni} = \mu_{ki} \sum_{n=1}^{N} r_{nk} \implies \mu_{ki} = \frac{\sum_{n=1}^{N} r_{nk} x_{ni}}{\sum_{n=1}^{N} r_{nk}}, \quad \forall k, i \ (1\text{pt})$$

This is the same type of update as that of GMM, but of course the responsibilities are *different* for the two models. As for GMM, this is part of the Maximization-step, where we update all the parameters and we keep the posterior fixed; the Mixture of Bernoullis 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).

e) The update rule computes a mean over the whole data, *i.e.* the 100 million pages. This is expensive and not necessary since we can approximate the mean well by using a much smaller sample size (1pt). A solution is to use Stochastic Gradient Descent. We have computed the gradient of the log-likelihood in point d), but note that the log-likelihood needs to be maximized. Stochastic gradient descent is used for minimizing an objective function, in this case the *negative* log-likelihood. The gradient with respect to μ_{ki} of the negative log-likelihood thus has an extra minus sign. Therefore for Stochastic Gradient *Descent* we have the update relative to page $\boldsymbol{x}_n$:

$$\begin{aligned}\mu_{ki}^{new} =& \mu_{ki} - \eta \frac{\partial}{\partial \mu_{ki}} (-\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})} [\ln p(X, Z|\boldsymbol{\pi}, \boldsymbol{\mu})]) \\ =& \mu_{ki} + \eta r_{nk} \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) \quad \forall_{k,i}. \text{ (1pt)}\end{aligned}$$

If Stochastic Gradient *Ascent* is chosen, the log-likelihood can be used as an objective function, leading to the update rule

$$\begin{aligned}\mu_{ki}^{new} =& \mu_{ki} + \eta \frac{\partial}{\partial \mu_{ki}} (\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})} [\ln p(X, Z|\boldsymbol{\pi}, \boldsymbol{\mu})]) \\ =& \mu_{ki} + \eta r_{nk} \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) \quad \forall_{k,i}.\end{aligned}$$

It is also OK to formulate the solution in terms of mini-batches. Let B be a random subset of pages $B \subset \{\boldsymbol{x}_n\}_{n=1}^N$. The mini-batch SGD mean update is:

$$\forall_{k,i} \quad \mu_{ki}^{new} = \mu_{ki} + \eta \frac{1}{|B|} \sum_{n:\boldsymbol{x}_n \in B} r_{nk} \left(\frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) .$$

Note that the factor $1/|B|$ is not necessary for points.

f) We could use PCA for dimensionality reduction, which retains most of the variance contained in the original data. PCA works as follows (2pt):

- Compute the sample mean: $\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^N \boldsymbol{x}_n$.
- Compute the sample covariance matrix: $\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^\top$.
- Compute the eigenvector decomposition of $\mathbf{S}$ up to the $H < D$ largest eigenvalue: $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top = \mathbf{S}$.
- Project the vectors into a H-dimensional representation: $\tilde{\boldsymbol{x}}_n = \mathbf{U}^\top (\boldsymbol{x}_n - \overline{\boldsymbol{x}})$.

Unfortunately, we will be losing the binary representation of the word vectors, therefore a Mixture of Bernoullis will not be appropriate (1pt). Notice that other algorithms for dimensionality reduction would be also fine as an answer.