Learning and Intelligent Systems

Final Exam

Aug 13, 2016

Time limit:	120 minutes
Number of pages:	19
Total points:	100

You can use the back of the pages if you run out of space. Collaboration on the exam is strictly forbidden. Please show *all* of your work and always *justify* your answers.

Please write your answers with a *pen*.

(1 point) Please fill in your student ID and full name (LASTNAME, FIRSTNAME) in capital letters.

Please leave the table below empty.

Problem	**Maximum points**	**Obtained**
1.	12	
2.	10	
3.	13	
4.	20	
5.	16	
6.	16	
7.	12	
Total	100	

1. Linear Regression (12 points)

Assume we have a data set of n points in d dimensions, represented by feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and corresponding output vector $\boldsymbol{y} \in \mathbb{R}^n$. Let $\boldsymbol{T} = \tau \boldsymbol{I}_d$, where $\tau \geq 0$, and $\boldsymbol{I}_d$ is the $d \times d$ identity matrix. We define a new feature matrix $\boldsymbol{X'} \in \mathbb{R}^{(n+d) \times d}$ and output vector $\boldsymbol{y'} \in \mathbb{R}^{n+d}$ as follows:

$$\boldsymbol{X}' = \begin{bmatrix} \boldsymbol{X} \\ \hline \boldsymbol{T} \end{bmatrix} = \left[\begin{array}{ccc} & \boldsymbol{X} & \\ \hline \tau & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tau \end{array}\right], \qquad \boldsymbol{y}' = \begin{bmatrix} \boldsymbol{y} \\ \hline \boldsymbol{0} \end{bmatrix}$$

(8 points) (i) Write down the least squares objective function for the new data set $(\boldsymbol{X}', \boldsymbol{y}')$ for a linear regression with weight vector $\boldsymbol{w} \in \mathbb{R}^d$. What is the difference from the objective for the original data set $(\boldsymbol{X}, \boldsymbol{y})$?

(4 points) (ii) What type of regularization do the d added data points impose? What is the role of τ?

2. Short questions (10 points)

(6 points) (i) Consider the two classification problems with two classes, which are illustrated by circles and crosses in the plots below. In each of the plots below, one of the following classification methods has been used, and the resulting decision boundary is shown:

A) Linear SVM

B) Kernelized SVM (Polynomial kernel of order 2)

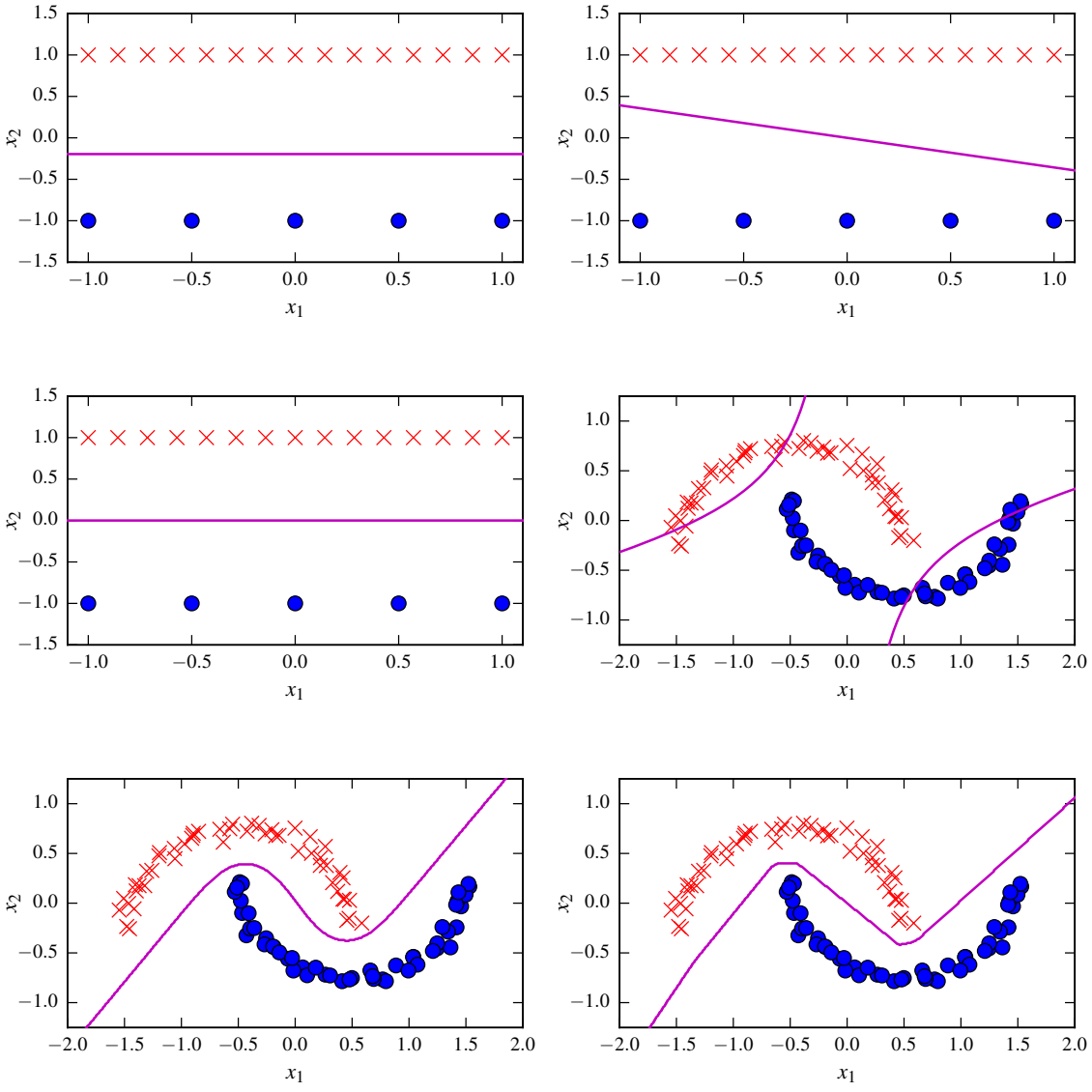
C) Perceptron

D) Logistic Regression

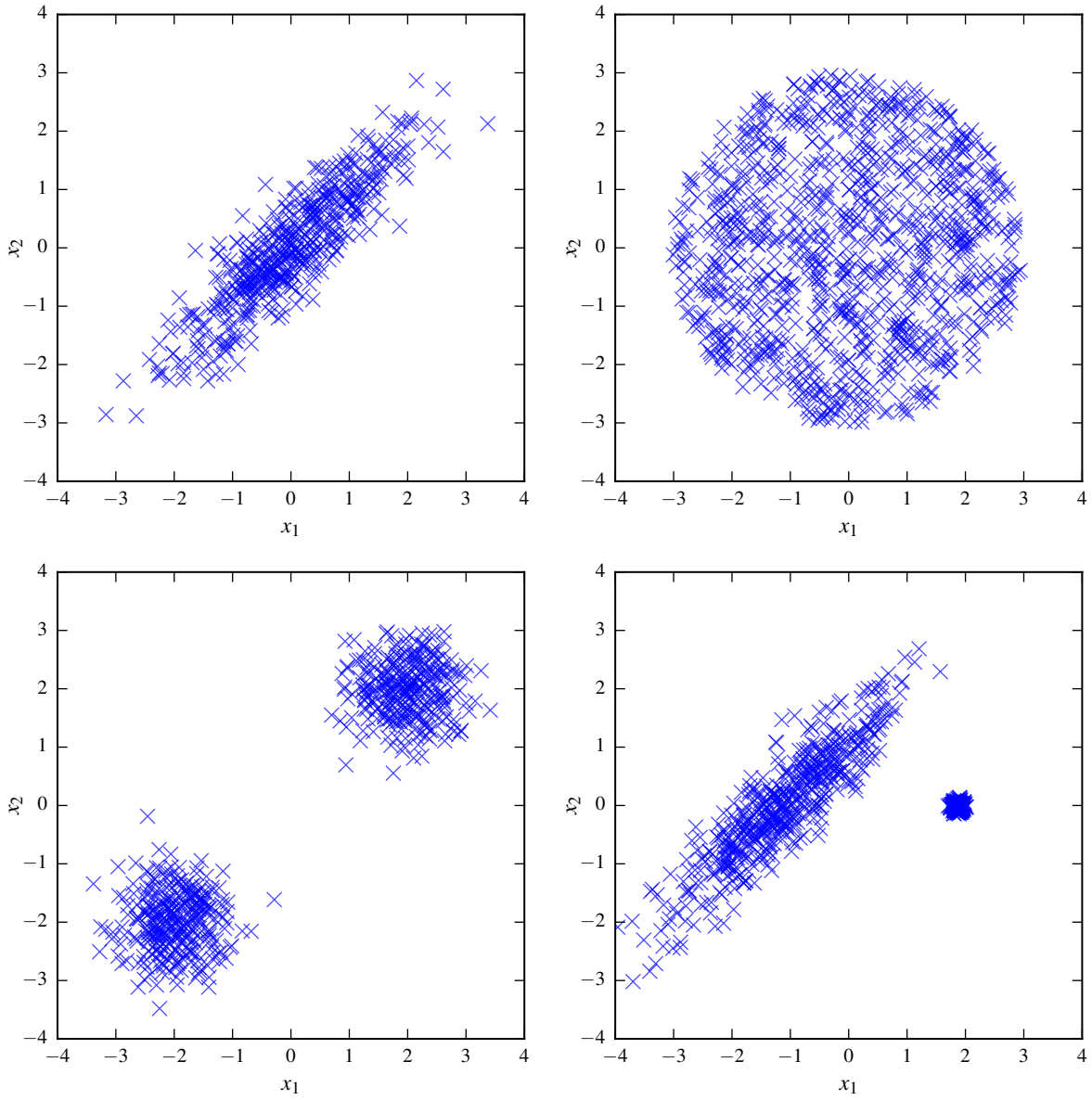
E) Neural Network (1 hidden layer with 10 rectified linear units)

F) Neural Network (1 hidden layer with 10 tanh units)

Assign each of the previous methods to exactly one of the following plots (in a one to one correspondence) by annotating the plots with the respective letters.



(4 points) (ii) In each figure below, draw the first principal component of the data. The data is centered. In the two figures at the bottom both clusters have the same number of data points.



3. Clustering (13 points)

Assume we have three one-dimensional data points $x_1 = 0, x_2 = 2, x_3 = 3$, which we want to cluster using the k-means algorithm. Note that, if at any iteration no point is assigned to some center μ_j, this center is not updated during that iteration.

(4 points) (i) For $k = 1$, compute the global minimizer of the k-means objective.

(4 points) (ii) For $k = 2$, if we initialize one cluster center as $\mu_1 = 0$ and the other as $\mu_2 = 10$, what will happen in the subsequent iterations of the k-means algorithm?

(5 points) (iii) For $k = 3$, what is a global minimizer of the k-means objective? Is there a local minimizer that is not global?

4. Naive Bayes Classifier (20 points)

Consider a naive Bayes classifier with a binary class $Y \in \{0, 1\}$ and three binary features $X_1, X_2, X_3 \in \{0, 1\}$. You are given a set D of n training examples, i.e.

$$D = \{(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, y^{(1)}), \ldots, (x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, y^{(n)})\},$$

where $x_1^{(k)}, x_2^{(k)}, x_3^{(k)}$ are the realizations of X_1, X_2 and X_3 in the k^{th} training example, respectively, and where $y^{(k)} \in \{0, 1\}$ is the class label of the k^{th} training example.

(4 points) (i) Write down the joint distribution $P(X_1, X_2, X_3, Y)$ of the naive Bayes classifier for the above setting. State how to compute the class posterior distribution $P(Y \mid X_1, X_2, X_3)$ from the joint distribution.

(4 points) (ii) Assume that we have trained the model from the given data D using maximum likelihood estimation. State the resulting estimated prior $P(Y)$ and the class-conditional distributions $P(X_i \mid Y)$. No proofs are necessary, you can just state the resulting distributions.

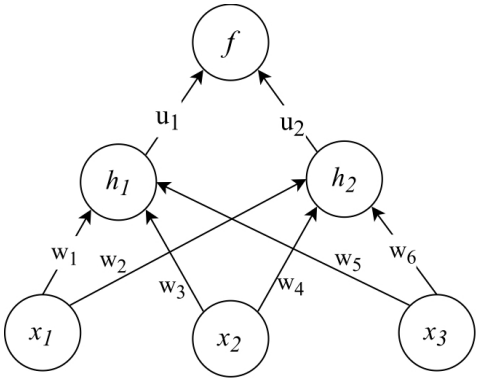
(6 points) (iii) Suppose for some example the posterior is $P(Y = 1 \mid X_1 = x_1, X_2 = x_2, X_3 = x_3) = p$ and assume that there is a cost c_{m} for misclassification (that is, classifying a sample with label 1 as having label 0 or vice versa). For each prediction you have the choice between trusting the naive Bayes classifier or asking a human expert with cost $c_{\mathrm{h}} < c_{\mathrm{m}}$, who is 100% accurate. Which of these two choices minimizes the expected cost as a function of p?

(6 points) (iv) We now consider a naive Bayes classifier with d binary features $X_1, \ldots, X_d \in \{0, 1\}$. We estimate $P(Y = 1) = P(Y = 0) = 0.5$, $P(X_i = 1|Y = 0) = \theta$ and $P(X_i = 1|Y = 1) = 1 - \theta$, where $\theta \in [0, 0.5)$ for all $i \in \{1, \ldots, d\}$. Given a new example with $X_1 = \ldots = X_d = 1$, determine the smallest number of features d such that the human expert is not consulted (the smallest number should be a function of θ, and the costs c_{h} and c_{m}).

5. Artificial Neural Networks (16 points)

Consider the following neural network with two logistic hidden units h_1, h_2, and three inputs x_1, x_2, x_3. The output neuron f is a linear unit, and we are using the squared error cost function $E = (y - f)^2$. The logistic function is defined as $\rho(x) = 1/\left(1 + e^{-x}\right)$.

[*Note: You can solve part (iii) without using the solution for part (ii).*]



(4 points) (i) Consider a single training example $\boldsymbol{x} = [x_1, x_2, x_3]$ with target output (label) y. Write down the sequence of calculations required to compute the squared error cost (called forward propagation).

(8 points) (ii) A way to reduce the number of parameters to avoid overfitting is to tie certain weights together, so that they share a parameter. Suppose we decide to tie the weights w_1 and w_4, so that $w_1 = w_4 = w_{\text{tied}}$. What is the derivative of the error E with respect to w_{tied}, i.e. $\nabla_{w_{\text{tied}}} E$?

(4 points) (iii) For a data set $D = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \cdots, (\boldsymbol{x}^{(n)}, y^{(n)})\}$ consisting of n labeled examples, augment the pseudocode of the stochastic gradient descent algorithm below with learning rate η_t for optimizing the weight w_{tied} (assume all the other parameters are fixed).

```
begin
    w_tied ← 0, η_t = 1/t;
    for t = 1 to T do
        // Fill in code to implement SGD
    end
end
```

6. Kernels (16 points)

Ω represents a ground set of n elements denoted as $\Omega = \{1, 2, \ldots, n\}$. Let $X \subseteq \Omega$ and $Y \subseteq \Omega$ denote any two subsets of Ω. In this question, we consider kernel functions over the space of subsets of Ω.

For each of the following functions, determine which are valid kernel functions and which are not. You can either provide a proof that it is a kernel, or provide a counterexample. In your answers, you may directly use the composition rules of kernels discussed in the lecture slides.

$|\cdot|$ denotes the cardinality of (i.e., number of elements contained in) a set.

(4 points) (i) $k(X, Y) = |X \setminus Y|$. Here, the operation $\setminus$ denotes the difference of two sets, i.e., the set of elements which are in X but not in Y.

(4 points) (ii) $k(X, Y) = |X \cap Y|$. Here, the operation $\cap$ denotes the intersection of two sets.

[*Hint: You might want to represent each* $X \subseteq \Omega$ *as a binary vector* $[x_1, x_2, \ldots, x_n]$*, where* $x_i = 1$ *if element* $i \in X$ *and* 0 *otherwise.*]

(4 points) (iii) $k(X, Y) = |X \cup Y|$. Here, the operation $\cup$ denotes the union of two sets.

(4 points) (iv) $k(X, Y) = 2^{|X \cap Y|} + 2^{|X \cup Y|}$.

7. Expectation Maximization (12 points)

Assume that we have a categorical distribution that represents drawing a red, green, or blue ball with probabilities $p_{\mathrm{r}} = 0.5, p_{\mathrm{g}} = \theta, p_{\mathrm{b}} = 0.5 - \theta$ respectively, where $\theta \in [0, 0.5]$ is an unknown parameter. After repeatedly and independently drawing n balls from this distribution, we know that the sum of red and green balls is $X_{\mathrm{r}} + X_{\mathrm{g}} = \alpha$, and the number of blue balls is $X_{\mathrm{b}} = n - \alpha := \beta$. We would like to use expectation maximization to estimate the value of θ.

If we knew the number of balls of each color to be x_{r}, x_{g}, x_{b}, we could compute the likelihood of our data set by noting that the counts follow a multinomial distribution,

$$P(X_{\mathrm{r}} = x_{\mathrm{r}}, X_{\mathrm{g}} = x_{\mathrm{g}}, X_{\mathrm{b}} = x_{\mathrm{b}} \mid \theta) = \frac{1}{Z} 0.5^{x_{\mathrm{r}}}\, \theta^{x_{\mathrm{g}}}\, (0.5 - \theta)^{x_{\mathrm{b}}},$$

and maximize this likelihood with respect to θ. Z is a normalization constant independent of θ.

However, in our problem we do not know the exact values of x_{r}, x_{g}, x_{b}, therefore we will instead iteratively maximize the *expected* log-likelihood of the data set given our observations α, β, and the previous estimate $\theta^{(k)}$.

If we define the conditional distribution

$$Q^{(k)}(X_{\mathrm{r}}, X_{\mathrm{g}}, X_{\mathrm{b}}) := \mathbb{P}\left[X_{\mathrm{r}}, X_{\mathrm{g}}, X_{\mathrm{b}} \;\middle|\; \alpha, \beta, \theta^{(k)}\right],$$

then the expected log-likelihood mentioned above can be written as

$$\mathcal{L}^{(k)}(\theta) := \mathbb{E}_{Q^{(k)}}\left[\log P(X_{\mathrm{r}}, X_{\mathrm{g}}, X_{\mathrm{b}} \mid \theta)\right].$$

[*Questions on the following pages*]

(4 points) (i) If we define $\xi_{\mathrm{r}}^{(k)} := \mathbb{E}_{Q^{(k)}}\left[X_{\mathrm{r}}\right]$ and $\xi_{\mathrm{g}}^{(k)} := \mathbb{E}_{Q^{(k)}}\left[X_{\mathrm{g}}\right]$, write $\mathcal{L}^{(k)}$ as a function of $\xi_{\mathrm{r}}^{(k)}$, $\xi_{\mathrm{g}}^{(k)}$, α, β and θ.

(4 points) (ii) Compute $\theta^{(k+1)}$ by maximizing $\mathcal{L}^{(k)}$ with respect to θ. (*M-step*)

(4 points) (iii) Compute $\xi_{\mathrm{g}}^{(k)}$ by noting that X_{g} follows a binomial distribution. (*E-step*)

[*Hint: The mean of a binomial distribution with* m *trials and success probability* p *is* pm.]