Learning and Intelligent Systems

Final Exam

Aug 10, 2017

Time limit:	120 minutes
Number of pages:	14
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.	17	
2.	15	
3.	14	
4.	21	
5.	17	
6.	15	
Total	100	

1. Linear Regression (17 points)

In this problem you will help Ada solve a linear regression problem. From the domain experts she has learned that it makes sense to use the following regularizer[1],

$$R(\mathbf{w}) = \sum_{i=1}^{d-1} |w_i - w_{i+1}|$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector. She is given n data points $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$, where each $\mathbf{x}_i \in \mathbb{R}^d$ and each $y_i \in \mathbb{R}$. Hence, she has to *minimize* the following objective

$$f(\mathbf{w}) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \underbrace{(\mathbf{w}^T \mathbf{x}_i - y_i)^2}_{\text{loss}(\mathbf{w}|y_i, \mathbf{x}_i)}}_{L(\mathbf{w})} + \lambda R(\mathbf{w}).$$

(6 points) (i) Ada wrote a program and then solved the above problem for the *same data points* and four *different* positive penalizers $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$. Unfortunately, she has misnamed the files holding the results and does not know which file corresponds to which λ_i. Your task is to help Ada by assigning to each file the corresponding λ_i that was used. Please justify your answer.

File name	Computed weight vector $\mathbf{w}^*$	Penalizer
solution_a.pkl	$(1, 1, 2, 2, 1, 1)$	
solution_b.pkl	$(9, 10, 10, 8, 2, 2)$	
solution_c.pkl	$(2, 2, 4, 5, 5, 5)$	
solution_d.pkl	$(1, 2, 2, 2, 3, 1)$	

[1] This regularizer makes sense if we would like to prefer solutions whose entries do not change much between adjacent coordinates.

(5 points) (ii) Ada's colleague Alan wrote another program to solve the same optimization problem, but arrived at a different optimum for the same penalizer $\lambda > 0$. Does this necessarilly means that one of them has an implementation bug?

(6 points) (iii) To ensure that her algorithm is correctly implemented, Ada wants to implement the following test procedure. First, come up with some synthetic distribution $P(\mathbf{x}, y)$ where the data comes from. Then, compute the optimal vector $\mathbf{w}^*$ on a finite sample from $P(\mathbf{x}, y)$, and finally compute the *generalization error* of $\mathbf{w}^*$. She defined the distribution generating the data as follows,

$$P(\mathbf{x}, y) = \begin{cases} \frac{1}{8} & \text{if } \mathbf{x} \in \{0,1\}^3 \text{ and } y = x_1 + 2x_2 + 2x_3, \\ 0 & \text{otherwise,} \end{cases}$$

and computed the vector $\mathbf{w}^* = (2, 2, 2)$ based on a finite sample. What is the *generalization error* of $\mathbf{w}^*$?

2. Classification Performance Measures (15 points)

We have a database of images consisting of 3 classes Cat, Dog and Bird and we train a classifier which takes an image as an input and assigns a class label as an output. When we run our classifier on our test set we obtain the following confusion matrix:

		Actual		
		Cat	Dog	Bird
Predicted	Cat	3	1	2
	Dog	0	4	3
	Bird	1	1	5

In class we have discussed several performance measures for binary classification. In order to answer the following 3 questions you need to generalize these measures to the above multiclass problem.

(1 point) (i) Calculate the accuracy of the classifier.

(2 points) (ii) Calculate the precision and recall of the classifier for class Cat.

(1 point) (iii) Write down a decision rule which always ensures a recall of 1 for class Dog.

The rest of this question discusses performance measures for the case of binary classification.

(4 points) (iv) You have several statements about precision-recall (PR) curve, receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) with a binary classifier which outputs probabilities for each class. For each of the statements below, decide whether they are true or false. (You get 1 point for a correct answer, 0 points for a blank answer, and -1 point for a wrong answer. You cannot get less than 0 points in total.)

a) ROC curve can be increasing or decreasing according to the performance of the classifier.
[ ] True [ ] False

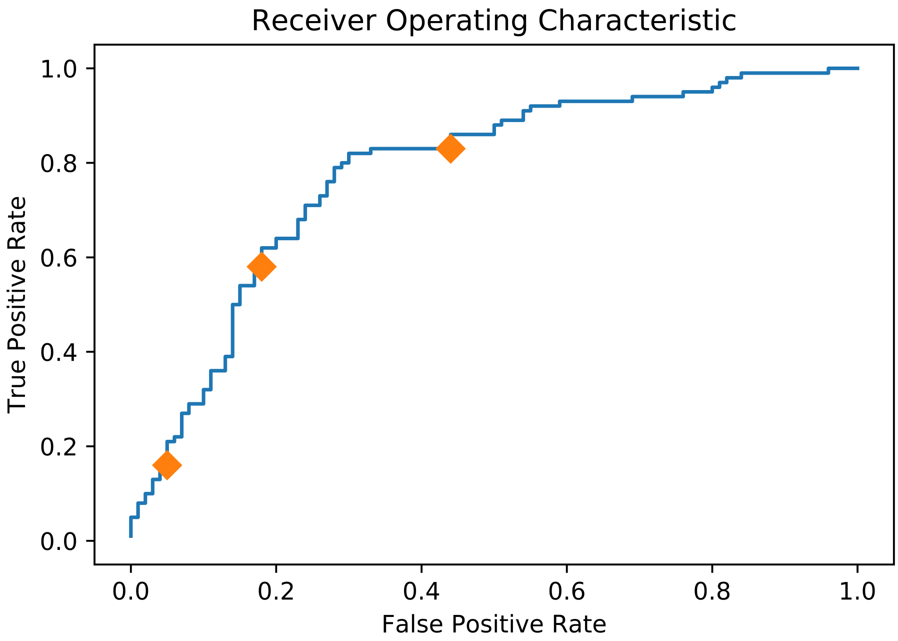
b) In PR curve, when recall decreases precision always increases.
[ ] True [ ] False

c) True positive rate and recall measure the same thing.
[ ] True [ ] False

d) AUC is always between 0 and 1.
[ ] True [ ] False

(5 points) (v) A scientist measures Globulin (a blood protein) level in patients. We know that the Globulin level (which we call G) in diseased people follows a normal distribution with mean 1 and variance 1. In healthy people, G follows a normal distribution with mean 0 and variance 1. After measuring G, the scientist classifies a person as diseased (positive) if $G > \lambda$, where $\lambda \in \mathbb{R}$ is a fixed threshold parameter. Write down the *expected* true positive rate, TPR(λ), in terms of the cumulative distribution function of standard normal random variable, $\Phi(x)$, and of λ.

(2 points) (vi) Now, suppose we have defined three thresholds: $\lambda_1 < \lambda_2 < \lambda_3$. Place them on top of the diamonds on the following ROC curve.



3. Markov Models (14 points)

Jimmy has recently taken up stock trading. To outfox his competitors he decided to model the daily movement of the swiss market index as a first-order Markov model, $(S_t)_{t\geq 1}$. According to this model, each day the index can stay roughly constant (C), move significantly up (U), or significantly down (D). Also he assumes that index movements are independent across different weeks. To learn the model he has gathered two weeks' worth of movement data that look as follows:

$$\mathcal{W} = \{(U, U, C, C, U), (C, D, D, C, D)\}.$$

(3 points) (i) How many independent parameters need to be learned to fully specify the model? Briefly describe what the parameters represent.

(4 points) (ii) What are the estimated model parameters according to the maximum likelihood estimator?

(4 points) (iii) According to the learned model, what is the probability of observing the sequence (U, D, U, U, U)? Do you see an issue here? Briefly describe a way to alleviate this issue.

(3 points) (iv) Given that $S_4 = U$, are S_5 and S_2 independent of each other? Prove your answer.

4. Poisson Naive Bayes (21 points)

In this task we will use the Naive Bayes model for binary classification. Let $\mathcal{Y} = \{0, 1\}$ be the set of labels and $\mathcal{X} = \mathbb{N}^d$ a d-dimensional features space ($\mathbb{N} = \{0, 1, 2, \dots\}$). You are given a training set $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ of n labeled examples $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$.

(1 point) (i) Is the Naive Bayes model a generative or a discriminative model? Justify your answer.

(4 points) (ii) Let λ be a positive scalar, and assume that $z_1, \dots, z_m \in \mathbb{N}$ are m iid observations of a λ-Poisson distributed random variable. Find the maximum likelihood estimator for λ in this model. *(Hint: A λ-Poisson distributed random variable Z takes values $k \in \mathbb{N}$ with probability $P(Z = k) = e^{-\lambda} \frac{\lambda^k}{k!}$.)*

(5 points) (iii) Let's train a Poisson Naive Bayes classifier using maximum likelihood estimation. Define appropriate parameters $p_0, p_1 \in [0, 1]$, and vectors $\lambda_0, \lambda_1 \in \mathbb{R}^d$, and write down the joint distribution $P(X, Y)$ of the resulting model. *(Note that the following should be satisfied for the parameters: $p_0 + p_1 = 1$, and λ_0, λ_1 are vectors with non-negative components.)*

(5 points) (iv) Now, we want to use our trained model from (iii) to minimize the misclassification probability of a new observation $\mathbf{x} \in \mathcal{X}$, i.e. $y_{\text{pred}} = \operatorname{argmax}_{y \in \mathcal{Y}} P(y|X = \mathbf{x})$. Show that the predicted label y_{pred} for $\mathbf{x}$ is determined by a hyperplane, i.e., that $y_{\text{pred}} = \left[\mathbf{a}^\top \mathbf{x} \geq b\right]$ for some $\mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}$.

The rest of this question concerns Bayesian decision theory (but it does not concern Naive Bayes).

(3 points) (v) Instead of simply predicting the most likely label, one can define a cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such that $c(y_{\text{pred}}, y_{\text{true}})$ is the cost of predicting y_{pred} given that the true label is y_{true}. Define the Bayes optimal decision rule for a cost function $c(\cdot, \cdot)$, with respect to a distribution $P(X, Y)$.

(3 points) (vi) Write down a cost function such that the corresponding decision rule that you have defined in (v) for this cost coincides with a decision rule that minimizes the misclassification probability, i.e., $y_{\text{pred}} = \operatorname{argmax}_{y \in \mathcal{Y}} P(y|X = \mathbf{x})$.

5. Kernels (17 points)

An SVM (support vector machine) enables to find a binary classifier based on a give a set of n binary labeled training points, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$; where each $\mathbf{x}_i \in \mathbb{R}^d$ and each $y_i \in \{0, 1\}$. There are however several options to train an SVM which give rise to possibly different classifiers. Concretely, we can choose to do one of the following,

- Directly train an SVM over the training set with the original features, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.
- Choose a feature map $\phi(\cdot)$, and train an SVM over the embedded training set, $\{(\phi(\mathbf{x}_1), y_1), \ldots, (\phi(\mathbf{x}_n), y_n)\}$.
- Choose a valid kernel $k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$, and solve the kernelized SVM problem.

Note: In the rest of this question when we write kernel we mean a valid kernel.

(9 points) (i) You have several statements about kernels and the use of kernels/feature-maps in training SVMs. For each of the statements below, decide whether they are true or false. (You get 1 point for a correct answer, 0 points for a blank answer, and -1 point for a wrong answer. You cannot get less than 0 points in total.)

a) For any kernel there exists an equivalent feature map.
[ ] True [ ] False

b) For any feature map there exists an equivalent kernel.
[ ] True [ ] False

c) Let $M \in \mathbb{R}^{d \times d}$ be a diagonal matrix with non-zero diagonal elements, and define:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top M \mathbf{x}', \ \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d,$$

then k is always a valid kernel.
[ ] True [ ] False

For the rest of this question recall the definitions of the linear kernel and the polynomial kernel of degree 2,

$$k_{\text{linear}}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}', \qquad k_{\text{poly}}(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + 1)^2, \qquad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$$

d) Using the polynomial kernel we are always ensured to obtain a lower *training* error compared to the linear kernel.
[ ] True [ ] False

e) Using the polynomial kernel we are always ensured to obtain a lower *generalization* error compared to the linear kernel.
[ ] True [ ] False

f) Using any kernel we are always ensured to obtain a lower *training* error compared to the linear kernel.
[ ] True [ ] False

g) Using any kernel we are always ensured to obtain a lower *generalization* error compared to the linear kernel.
[ ] True [ ] False

h) For any kernel, the optimal solution to the kernelized SVM problem can always be written as a linear combination of the training points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

[ ] True [ ] False

i) The optimal solution to the original problem (without kernel trick or feature map) is the same as the optimal solution we would get using the linear kernel.

[ ] True [ ] False

The rest of this question concerns specific kernels and feature maps.

(4 points) (i) For $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, and $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^2$, find a feature map $\phi(\mathbf{x})$, such that $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$.

(4 points) (ii) For the dataset $X = \{\mathbf{x}_i\}_{i=1,2} = \{(-3, 4), (1, 0)\}$ and the feature map $\phi(\mathbf{x}) = [x^{(1)}, x^{(2)}, \|\mathbf{x}\|]$, calculate the **Gram matrix** (for a vector $\mathbf{x} \in \mathbb{R}^2$ we denote by $x^{(1)}, x^{(2)}$ its components).

6. Mixture Models and Expectation-Maximization Algorithm (15 points)

Consider a one-dimensional Gaussian Mixture Model with 2 clusters and parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)$. Here (w_1, w_2) are the mixing weights, and (μ_1, σ_1^2), (μ_2, σ_2^2), are the centers and variances of the clusters. We are given a dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \subset \mathbb{R}$, and apply the EM-algorithm to find the parameters of the Gaussian mixture model.

(3 points) (i) Write down the complete log-likelihood that is being optimized, *for this problem.*

Assume that the dataset $\mathcal{D}$ consists of the following three points, $\mathbf{x}_1 = 1, \mathbf{x}_2 = 10, \mathbf{x}_3 = 20$. At some step in the EM-algorithm, we compute the expectation step which results in the following matrix:

$$R = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

where r_{ic} denotes the probability of $\mathbf{x}_i$ belonging to cluster c.
In the next questions, leave all results unsimplified, i.e. in fractional form.

(3 points) (ii) Given the above R for the expectation step, write the result of the maximization step for the mixing weights w_1, w_2. You can use the equations for maximum likelihood updates without proof.

(3 points) (iii) Do the same for μ_1, μ_2. Given the above R for the expectation step, write the result of the maximization step for the centers μ_1, μ_2 . You can use the equations for maximum likelihood updates without proof.

(3 points) (iv) Do the same for σ_1^2, σ_2^2. Given the above R for the expectation step, write the result of the maximization step for the variance values σ_1^2, σ_2^2. You can use the equations for maximum likelihood updates without proof.

(3 points) (v) The previous two questions are doing soft-EM. Calculate the maximization step of μ_1, μ_2 for hard-EM.