Introduction to learning and analysis of big data
Exercise 3

Dr. Sivan Sabato

Fall 2017/8

Submission guidelines, **please read and follow carefully**:

- You may submit the exercise in pairs.
- Submit using the submission system.
- The submission should be single file named "answers.pdf".
- For questions use the course Forum, or if they are not of public interest, send them via the course requests system.

Question 1. Implement the ridge-regression algorithm. **no need to submit your code**. Run the algorithm on the dataset `regdata.mat` provided on the course web page. Run the regression using $\lambda \in \{0, 1, 2, \ldots, 30\}$ on the training set X, Y provided in the data file. Try training-set sizes between 10 and 100. For each training set size that you try, find the value of λ that obtains the smallest mean-squared-error (the average squared loss) on the test set provided in the data file.

(a) (10pts) Submit a plot of the value of λ that minimizes the mean squared error on the test set as a function of the training set size m.

(b) (10pts) What trend to you expect in the plot based on what we learned in class? Explain.

(c) (5pts) Did you get this trend in the plot you submitted? If there are any differences, explain why they could occur.

(d) (15pts) In this data set, the label y of each example x was generated by setting $y = \langle w, x\rangle + \eta$, where w is a fixed vector which is the same for all examples in the data set, and η is a standard Gaussian random variable, $\eta \sim N(0, \sigma)$ for some $\sigma > 0$. η is drawn independently for each example in the data set. What is the Bayes-optimal predictor for this problem with respect to the squared loss? And how about the absolute loss? Prove your claims.

Question 2. (10pts) We saw in class that the LASSO regression algorithm minimizes the following objective function:

$$\text{Minimize}_{w\in\mathbb{R}^d} \lambda\|w\|_1 + \sum_{i=1}^{m} (\langle w, x_i\rangle - y)^2.$$

Write a quadratic program that solves this minimization program.

Hint: you will need to use auxiliary variables to represent $\|w\|_1$, similarly to what we showed in class for soft SVM.

Question 3. Let $\mathcal{X}$ be the set of all the undirected graphs over n vertices and let $\mathcal{Y} = \{0, 1\}$.

For a graph $x \in \mathcal{X}$, define the mapping $g : \mathcal{X} \to \mathbb{N}^n$, where coordinate i in the vector $g(x)$ is the number of i-cliques in the graph x.

Let $\mathcal{H}$ be a hypothesis class which includes all the functions $h : \mathcal{X} \to \mathcal{Y}$ that have the form: $h(x) \equiv \mathbb{I}[g(x) = v]$, for some vector $v \in \mathbb{N}^n$.

Suppose that $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$, and suppose that in this distribution, the label of a graph is completely determined by the numbers of i-cliques that the graph has for the various i.

(a) (10pts) Use the PAC-learning upper bounds that we showed in class to show that the sample complexity of learning $\mathcal{H}$ is $O(n^2 \log(n))$.

(b) (10pts) Adam and Ronnie want to learn a classifier that is guaranteed a low error with a probability of at least 95%. Ronnie used a training set of size m drawn from $\mathcal{D}$ to learn a classifier from $\mathcal{H}$. How many examples should Adam feed his classifier, so that his classifier has an error guarantee which is half the error guarantee of Ronnie's? Explain.

Question 4. Consider the following optimization objective for a regression problem, where $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$:

$$\text{Minimize}_{w \in \mathbb{R}^d} \lambda \|w\|_2^4 + \sum_{i=1}^{m} \exp^{|\langle w, x_i \rangle - y_i|}$$

(a) (7pts) Does the Representer Theorem apply for this objective function? Prove your claim.

(b) (13pts) Suppose we have a feature map $\psi : \mathcal{X} \to \mathbb{R}^n$, and a kernel function $K(x, x') := \langle \psi(x), \psi(x') \rangle$, and want to find a non-linear predictor $w \in \mathbb{R}^n$ that minimizes the objective function above in the feature space. In other words, we want to solve the following optimization problem:

$$\text{Minimize}_{w \in \mathbb{R}^n} \lambda \|w\|_2^4 + \sum_{i=1}^{m} \exp^{|\langle w, \psi(x_i) \rangle - y_i|}$$

Show how this can be done without directly representing any vectors in $\mathbb{R}^n$, by reformulating the problem to use the kernel function K. Prove the equivalence of the two problems.

Question 5. (10pts) Consider the following loss function for regression:

$$\ell(y, y') = (\ln(y) - \ln(y'))^2.$$

Calculate the Bayes-optimal predictor for this loss, $\text{argmin}_{h:\mathcal{X} \to \mathcal{Y}} \ell(h, \mathcal{D})$. Prove your claim using the definition of a Bayes-optimal predictor.