Introduction to Learning and Analysis of Big Data HW 4

Tomer Elgavish, Shay Kricheli

February 2020

Question 1

In this question we are asked to implement the ridge-regression algorithm which can be considered as regularized least squares - in which the cost functions penalizes the norm of w. Let $S = \{(x_i, y_i) \mid 1 \leq i \leq m\}$ be the training sample set where $x_i = (x_{1i}, x_{2i}, ..., x_{di})^T \in \mathbb{R}^d, y_i \in \mathbb{R};\ \forall 1 \leq i \leq m$. The optimization problem is as follows:

$$\min_{w \in \mathbb{R}^d} \left[\lambda ||w||^2 + \sum_{i=1}^{m} \left(\langle w, x_i \rangle - y_i \right)^2 \right]$$

Let us define over S:

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dm} \end{bmatrix} \in \mathbb{R}^{d \times m} \ ; \ Y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m$$

Thus the optimization problem can be written as follows:

$$\min_{w \in \mathbb{R}^d} \left[\lambda ||w||^2 + ||X^T w - Y||^2 \right]$$

As seen in class, this optimization objective can be considered a function of w which can be differentiated by the vector w using vector calculus and then equated to zero to get the optimal solution:

$$w^* = (XX^T + \lambda I_{d \times d})^{-1} XY$$

As defined in class, our predictor will be:

$$h_{w^*}(x) := \langle w^*, x \rangle$$

(a)

In this section we are asked to display a plot of λ^* (the value of λ that minimizes the mean squared error on the test set) as a function of m - the training set size, for $\lambda \in \{0, 1, ..., 30\}$ and $m \in \{10k \mid 1 \leq k \leq 10\}$. The plot is displayed below:

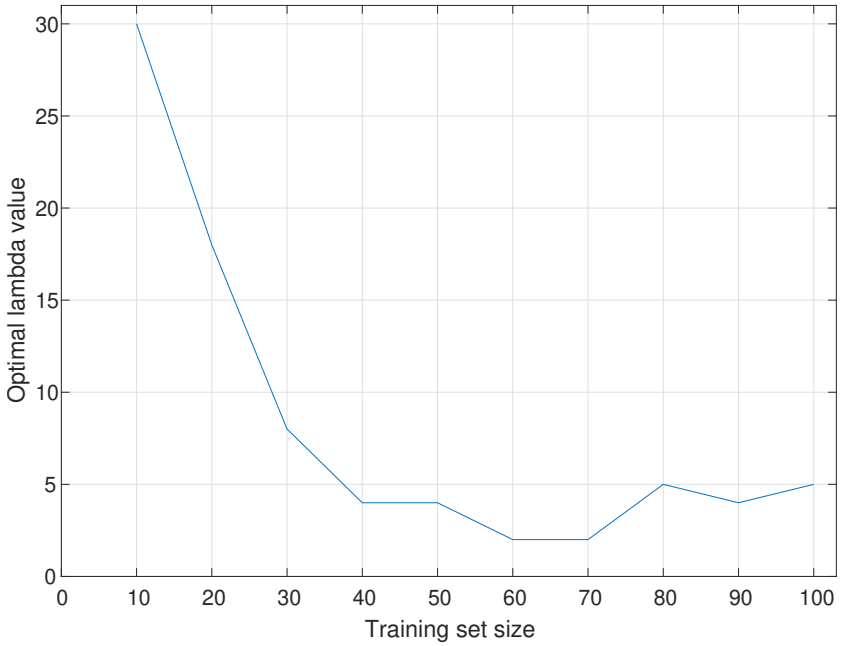

Figure 1: Q1 - Optimal λ value vs. Training Sample Size

(b)

In this section we are asked about the expected trend in the plot according to what was learned in class. In class we learned that:

$$\ell(h, \mathcal{D}) \leq \ell(h, S) + O\Big(\frac{B^2 R^2}{\sqrt{m}}\Big)$$

where: $R = \max_{i \leq m} ||x_i||, B = \max_{w: h_w \in \mathcal{H}} ||w_i||$. That means that the loss (as defined in class) on the distribution is bounded from above by the loss on the sample set plus a term of the form $\psi = O\big(\frac{B^2 R^2}{\sqrt{m}}\big)$. Thus the error on the test set - which is induced by the loss on the distribution - has a lower upper bound when m increases. When m is small, in order to decrease ψ, we'd want B to be small - which would happen when λ is bigger. Otherwise, when m is large, B can also be large so λ can be smaller in order to achieve the same value for the minimization problem. Thus λ^* - the optimal value for λ can also be smaller. In conclusion - as m increases we'd expect λ^* to decrease.

(c)

As can be seen - the expected trend occurs in the plot - as λ^* decreases as m increases.

(d)

As we learned in class - the Bayes optimal regressor that minimizes the squared loss of the form:

$$\mathbb{E}_{(X,Y)\sim D}\big[(h(X) - Y)^2\big]$$

over all $h : \mathcal{X} \to \mathcal{Y}$ is:

$$h^*(x) = \mathbb{E}_{(X,Y)\sim D}[Y|X = x]$$

Let us calculate it:

$$h^*(x) = \mathbb{E}_{(X,Y)\sim D}[Y|X = x] =$$
$$\mathbb{E}_{(X,Y)\sim D}[\langle w, x\rangle + \eta|X = x] \stackrel{1}{=}$$
$$\mathbb{E}_{(X,Y)\sim D}[\langle w, x\rangle|X = x] + \mathbb{E}_{(X,Y)\sim D}[\eta|X = x] \stackrel{2}{=}$$
$$\langle w, x\rangle + \mathbb{E}[\eta] \stackrel{3}{=} \langle w, x\rangle + 0 = \langle w, x\rangle$$

Let us explain each transition:

1. Linearity of the expected value operator
2. Given that $X = x$ and w is constant - $\langle w, x\rangle$ is constant and η is independant of X
3. The expected value of the Gaussian random variable η is 0

For the absolute loss - as we learned in class - the Bayes optimal regressor that minimizes the absolute loss of the form:

$$\mathbb{E}_{(X,Y)\sim D}\big[|h(X) - Y|\big]$$

over all $h : \mathcal{X} \to \mathcal{Y}$ is:

$$h^*(x) = MEDIAN_{(X,Y)\sim D}[Y|X = x]$$

Let us calculate it:

$$h^*(x) = MEDIAN_{(X,Y)\sim D}[Y|X = x] =$$
$$MEDIAN_{(X,Y)\sim D}[\langle w, x\rangle + \eta|X = x] \stackrel{1}{=}$$
$$\langle w, x\rangle + MEDIAN_{(X,Y)\sim D)}[\eta|X = x] \stackrel{2}{=}$$
$$\langle w, x\rangle + MEDIAN[\eta] \stackrel{3}{=}$$
$$\langle w, x\rangle + \mathbb{E}[\eta] = \langle w, x\rangle + 0 = \langle w, x\rangle$$

Let us explain each transition:

1. Given $X = x$, and since w is a fixed vector - we can infer that the inner product $\langle w, x\rangle$ is a constant. By the properties of a median, we can pull out the constant.
2. η is independant of X.
3. The median of a Gaussian random variable is it's expected value.

Question 2

(a)

First, we can see that the maximal size of a decision tree with depth of at most n is the size of a perfect binary tree with depth of n. The number of nodes in a perfect binary tree with depth of n is: $2^{n+1} - 1$. In addition, we can see that each node has $3d + 2$ options: if the node is a test attribute, it has $3d$ options, since for each feature there are 3 values of θ to be chosen from. If the node is a leaf, it has 2 possible values. Therefore, in conclusion, we can bound the size of our hypothesis class by:

$$|\bar{\mathcal{H}}_n| \leq (3d + 2)^{2^{n+1}-1} < (3d + 2)^{2^{n+1}}$$
$$\to |\bar{\mathcal{H}}_n| \leq (3d + 2)^{2^{n+1}} \blacksquare$$

(b)

We learned in class that this inequality is guaranteed for any $\delta, \varepsilon \in (0,1)$ from the PAC learning theorem only if the algorithm used is an ERM algorithm. We also saw in class that the ID3 algorithm (even with pruning) is not an ERM algorithm and thus it is not guaranteed and so Danny is wrong.

Question 3

(a)

The Naive-Bayes assumption does not hold for this distribution! Let us present a counter example: Let us look at the example $(-1,1)$ with the label -1. The Naive-Bayes assumption holds if: $\mathbb{P}[X = x|Y = y] = \Pi_{i=1}^{n}\mathbb{P}[X(i) = x(i)|Y = y]$. We will calculate both sides separately:

$$\mathbb{P}[X = x|Y = y] = \mathbb{P}[X = (-1,1)|Y = -1] \overset{1}{=} 0$$

The transition 1 is from the distribution's probabilities of the example: $(x, y) = ((-1,1), -1)$.

On the other hand:

$$\Pi_{i=1}^{n}\mathbb{P}[X(i) = x(i)|Y = y] = \mathbb{P}[X(1) = -1|Y = -1] \cdot \mathbb{P}[X(2) = 1|Y = -1] \overset{1}{=} \frac{1}{4} \cdot \frac{1}{5} = \frac{1}{20} \neq 0$$

Let us explain transition 1:

$$\mathbb{P}[X(1) = -1|Y = -1] = \frac{\frac{5}{60} + \frac{0}{60}}{\frac{20}{60}} = \frac{5}{20} = \frac{1}{4}$$

$$\mathbb{P}[X(2) = 1|Y = -1] = \frac{\frac{0}{60} + \frac{4}{60}}{\frac{20}{60}} = \frac{4}{20} = \frac{1}{5}$$

Thus we can conclude that the Naive-Bayes assumption doesn't hold.

(b)

As we learned in class, the Naive-Bayes algorithm outputs the following predictor:

$$h^*(x) = sign\Big(\sum_{i=0}^{n} \log\big(\frac{p_i}{1-p_i}\big) \cdot x(i)\Big)$$

We can see that we are in the symmetric case - since:

$$p_1 = \frac{\frac{2}{60} + \frac{8}{60}}{\frac{40}{60}} = \frac{10}{40} = \frac{1}{4} \; ; \; p_1' = \frac{\frac{5}{60} + \frac{0}{60}}{\frac{20}{60}} = \frac{5}{20} = \frac{1}{4}$$

$$p_2 = \frac{\frac{24}{60} + \frac{8}{60}}{\frac{40}{60}} = \frac{32}{40} = \frac{4}{5} \; ; \; p_2' = \frac{\frac{5}{60} + \frac{11}{60}}{\frac{20}{60}} = \frac{16}{20} = \frac{4}{5}$$

Also, $p_0 = \frac{40}{60} = \frac{2}{3}$.

We will now plug-in the values of $p_i's$ and get:

$$h^*(x) = sign\Big(\sum_{i=0}^{n} \log\big(\frac{p_i}{1-p_i}\big) \cdot x(i)\Big) =$$

$$sign\Big(\log 2 \cdot 1 + \log\frac{1}{3} \cdot x(1) + \log 4 \cdot x(2)\Big) =$$

$$sign\Big(\langle \hat{w}, \hat{x} \rangle\Big)$$

Where $\hat{w} = \left(\log 2, \log \frac{1}{3}, \log 4\right)$ and $\hat{x} = \big(1, x(1), x(2)\big)$.

Question 4

(a)

We saw in class that if we perform PCA to reduce the dimension of our data from d to k - then the distortion will be the sum of the $d-k$ lowest eigenvalues of $A = \sum_{i=1}^{m} x_i x_i^T = X^T X \in \mathbb{R}^{d \times d}$, where:

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} \end{bmatrix} \in \mathbb{R}^{m \times d}$$

We will now show that there are at least two eigenvalues of A that are equal 0. In algebra we learned the following theorems:

1. For any matrix B we have: $rank(B) = rank(B^T B)$.
2. The number of non-zero eigenvalues of any matrix B is at most $rank(B)$.
3. A matrix B is positive semi-definite $\iff$ each eigenvalue λ_i of B holds $\lambda_i \geq 0$.

Since there's a linear dependence between $x_t(3), x_t(4)$ and $x_t(1), x_t(2)$ - we have that $rank(X) = 2$. By theorem 1 we have that $rank(A) = rank(X^T X) = rank(X) = 2$. By theorem 2 - since $rank(A) = 2$, we have that there are at least $dim(A) - rank(A) = 4 - 2 = 2$ eigenvalues of A that are zero. ■
We will prove that A is positive semi-definite. Let $z \in \mathbb{R}^d$.

$$z^T A z = z^T X^T X z = (Xz)^T (Xz) = z'^T z' = ||z'||^2 \geq 0 \ \blacksquare$$

By theorem 3 - we have that each eigenvalue λ_i of A holds $\lambda_i \geq 0$. Thus both zero eigenvalues are the lowest - and thus the distortion is their sum which is 0.

(b)

We will show an example of experiment results that satisfy these equations such that the distortion of the PCA is larger than zero: Let us take $m = 4$ and the following examples:

$$x_1 = (1, 1, 2, 1) \ ; \ x_2 = (0, 1, 1, 1) \ ; \ x_3 = (2, 2, 12, 100) \ ; \ x_4 = (1, 2, 9, 64)$$

Let us now calculate the matrix A:

$$A = X^T X = \begin{bmatrix} 1 & 0 & 2 & 1 \\ 1 & 1 & 2 & 2 \\ 2 & 1 & 12 & 9 \\ 1 & 1 & 100 & 64 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 \\ 2 & 2 & 12 & 100 \\ 1 & 2 & 9 & 64 \end{bmatrix} = \begin{bmatrix} 6 & 7 & 35 & 265 \\ 7 & 10 & 45 & 330 \\ 35 & 45 & 230 & 1779 \\ 265 & 330 & 1779 & 14098 \end{bmatrix}$$

Since we reduce the dimension of the examples from 4 to 2, the distortion of the PCA will equal to the sum of the 2 lowest eigenvalues of the matrix A. By calculating the eigenvalues of A, we get: $\lambda_1 = 0.11$; $\lambda_2 = 0.63$; $\lambda_3 = 7.93$; $\lambda_4 = 14335.30$. Therefore, the distortion of the PCA is: $0.11 + 0.63 = 0.74 > 0$.

Question 5

In this question we are to claim whether the algorithm that minimizes the k-means objective learned in class satisfies each of the three specified axioms and prove our claims.

(a - Scale Invariance)

We will show that this axiom holds. Let $\mathcal{X}$ be some space, let $S \subseteq \mathcal{X}$ and let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be some metric. Let us denote by $C = \{C_1, C_2, ..., C_k\}$ the clustering partition returned by $F(S, \rho)$. Let $\alpha > 0 \in \mathbb{R}$ and let us denote by $C' = \{C'_1, C'_2, ..., C'_k\}$ the clustering partition returned by $F(S, \alpha\rho)$. Let us assume towards contradiction that $C \neq C'$. Thus, assuming the optimal solution is singular, we get:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \alpha\rho(x, \mu_i)^2 > \sum_{i=1}^{k} \sum_{x \in C'_i} \alpha\rho(x, \mu_i)^2$$

We can divide both sides of the inequality by α and get:

$$\sum_{i=1}^{k} \sum_{x \in C_i} \rho(x, \mu_i)^2 > \sum_{i=1}^{k} \sum_{x \in C'_i} \rho(x, \mu_i)^2$$

i.e. C' is a better solution than C - in contradiction to the fact that C is the optimal solution returned from $F(S, \rho)$. Therefore, in conclusion, $F(S, \rho) = F(S, \alpha\rho)$ ■.

(b - Richness)

We will show that this axiom doesn't hold. Let $\mathcal{X}$ be some space and let $S \subseteq \mathcal{X}$. For any valid partition C that we choose of S of size bigger than k (the k value from the k-means algorithm) - such as $k + 1$, we'll get $F(S, \rho) \neq C$ for all possible metrics ρ - since F returns a partition of size k ■.

(c - Consistency)

Let us consider the following example: Let $k = 2$ and let there be only 5 points $\mathcal{X} = \{x_1, x_2, x_3, x_4, x_5\}$. Let $S = \mathcal{X}$ and let us define:

$$\rho(x_i, x_j) = 1 \ ; \ \forall i, j \leq 4 \text{ s.t. } i \neq j$$
$$\rho(x_i, x_5) = 1 + \varepsilon \ ; \ \forall i \leq 4 \ (1 \gg \varepsilon > 0)$$

Since $k = 2$, there will be two clusters, that one can see that according to the k-means algorithm and the metric defined will be the following:

$$C_1 = \{x_1, x_2, x_3, x_4\} \ ; \ C_2 = \{x_5\}$$

Let us assume:

$$\mu(C_1) = x_1 \ ; \ \mu(C_2) = x_5$$

Thus the returned output of the algorithm will be:

$$F(S, \rho) = C = (C_1, C_2)$$

Now let us consider another metric:

$$\rho'(x_1, x_2) = \rho'(x_3, x_4) = \alpha \ (1 \gg \alpha > \varepsilon > 0)$$
$$\rho'(x_i, x_j) = \rho(x_i, x_j) \ ; \ \forall i, j \text{ s.t. } \neg\big((i = 1 \wedge j = 2) \vee (i = 3 \wedge j = 4)\big)$$

The only changes were to $(x_1, x_2), (x_3, x_4)$ which are all in the same cluster, such that:

$$\rho'(x_1, x_2) \leq \rho(x_1, x_2)$$
$$\rho'(x_3, x_4) \leq \rho(x_3, x_4)$$

Let us assume towards contradiction that the algorithm now returns the same clusters C (with the same centroids). Then the 2-means objective value will be:

$$G_{2-\text{means}}(C_1, C_2) = \min_{\{\mu_i\}_{i=1}^2} \sum_{i=1}^{2} \sum_{x \in C_i} \rho(x, \mu_i)^2 =$$
$$\rho(x_1, x_1)^2 + \rho(x_2, x_1)^2 + \rho(x_3, x_1)^2 + \rho(x_4, x_1)^2 + \rho(x_5, x_5)^2 =$$
$$0 + \alpha^2 + 1 + 1 + 0 = 2 + \alpha^2$$

Let us define the following different clusters:

$$C_1' = \{x_1, x_2\} \ ; \ C_2' = \{x_3, x_4, x_5\}$$

Let us assume:

$$\mu(C_1') = x_1 \ ; \ \mu(C_2') = x_3$$

Now the 2-means objective value will be:

$$G_{2-\text{means}}(C_1', C_2') =$$
$$\rho'(x_1, x_1)^2 + \rho'(x_2, x_1)^2 + \rho'(x_3, x_3)^2 + \rho'(x_4, x_3)^2 + \rho'(x_5, x_3)^2 = 0 + \alpha^2 + 0 + \alpha^2 + (1 + \varepsilon)^2 =$$
$$1 + 2(\alpha^2 + \varepsilon) + \varepsilon^2$$

For $\varepsilon = \alpha = 0.1$:

$$G_{2-\text{means}}(C_1', C_2') = 1 + 2(0.1^2 + 0.1) + 0.1^2 = 1.23 < 2 + \alpha^2 = G_{2-\text{means}}(C_1, C_2)$$

In contradiction. Thus we'll get:

$$F(S, \rho') \neq (C_1, C_2) = F(S, \rho)$$

and thus the k-means algorithm does not satisfy consistency ■.