Homework 3

Due: March 13, 2018 at 7:00PM

Written Questions

Problem 1

(10 points)
Recall the update process for K-means. Let C_i denote the cluster of observations from sample S nearest to mean μ_i. Let t denote the iteration within the K-means update process. (*Note*: This is just an iterative restatement of the update rule from the slides):

$$\forall i \in \{1, ..., k\} : C_i^{t+1} = \{x \in S : i = \underset{j}{\operatorname{argmin}} ||x - \mu_j^t||\}$$

$$\forall i \in \{1, ..., k\} : \mu_i^{t+1} = \frac{1}{|C_i^{t+1}|} \sum_{x \in C_i^{t+1}} x$$

The loss for a K-means classifier is defined as follows:

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Prove that the choices of μ_i^t minimizes loss given clusters $C_1^t, ..., C_k^t$.

Solution

Consider each cluster C_i^t. We must select μ_i^t that minimizes

$$\mathcal{L} = \sum_{x \in C_i^t} ||x - \mu_i^t||^2$$

Using the fact that $||x||^2 = x^T x$, we can expand this loss function:

$$= \sum_{x \in C_i^t} (x - \mu_i^t)^T (x - \mu_i^t)$$

$$= \sum_{x \in C_i^t} (x^T x - 2x^T \mu_i^t + (\mu_i^t)^T \mu_i^t)$$

Differentiating with respect to μ_i^t yields:

$$\frac{d\mathcal{L}}{d\mu_i^t} = \sum_{x \in C_i^t} (-2x + 2\mu_i^t)$$

Setting the derivative to zero and solving, we obtain

$$\sum_{x \in C_i^t} (-2x + 2\mu_i^t) = 0$$

$$|C_i^t| \mu_i^t = \sum_{x \in C_i^t} x$$

$$\mu_i^t = \frac{\sum_{x \in C_i^t} x}{|C_i^t|}$$

This shows that our choice of μ_i^t is indeed a local extrema. Our loss function is convex (may be verified by showing the second derivative is 2, which is always positive), so this extrema must be a minimum.

Problem 2

(8 points)
Suppose we have a weighted coin that flips heads with probability p. We would like to predict the outcome of the next coin flip.

a. Suppose p is known. If $p \geq \frac{1}{2}$, we will predict heads for the next outcome as we know this to be the most likely possibility. If $p < \frac{1}{2}$, we will predict tails. What is the expected error of a single coin flip? *Note:* This method provides an optimal prediction for the next coin flip.

b. Suppose p is unknown. We instead opt to flip the coin, and use the result of that coin flip as a prediction for the following coin flip. What's the expected error of this prediction method? *Note:* This prediction method may be seen as analogous to a 1-NN model. We are making a prediction for an unknown label based upon the label of the nearest neighbor: our instanced coin flip.

c. Let E_a and E_b denote the expected error of our prediction methods in parts a and b respectively. Find c such that $E_b \leq cE_a \ \forall p \in [0, 1]$.

d. What does this analysis tell us about the best performance we can hope for in a 1-NN model (under the formal model discussed in lecture - see slide 9)?

Solution

a. $\min(p, 1-p)$

If $p \geq 0.5$, we will guess heads so our expected error is $1-p$. Alternatively, if $p < 0.5$, we will guess tails, so the expected error will be p. Therefore, our expected error is $\min(p, 1-p)$.

b. $2p(1-p)$

You can see this by enumerating all possible outcomes. There are four possible outcomes: HH, HT, TH, TT. In the case of TH, HT, our error is 1, otherwise it is zero. HT and TH each occur with probability $p(1-p)$.

c. $c = 2$

$$E_b \leq cE_a$$
$$2p(1-p) \leq c \min(p, 1-p)$$

Consider the case where $p < \frac{1}{2}, p \neq 0$ (We let $p \neq 0$, so that we can divide by p. We can trivially see that when $p = 0$, $E_a = E_b = 0$ so we can ignore this case).

$$2p(1-p) \leq cp$$
$$2(1-p) \leq c$$

Notice that since $p < \frac{1}{2}$, $1 - p \in (\frac{1}{2}, 1]$

The LHS is maximized in the limit as $p \to 0$ (Remember $p \neq 0$). $\lim_{p \to 0} 2(1-p) = 2$. Therefore $c = 2$.

Note: We can obtain the same result by considering the case in which $p > \frac{1}{2}, p \neq 1$, by symmetry.

d. Even under ideal circumstances, Nearest Neighbor can be as bad as 2 $\times$ optimal because it is getting its label from a neighbor's random label (probability matching).

Problem 3

(10 points)
For any hypothesis space H, to show that the VC Dimension of H is d, you should show each of the following:

- There exists a set C of size d that is shattered by H
- There exists no set C of size $d+1$ that is shattered by H

1. Compute the VC dimension for the following hypothesis spaces:

   a. The class of signed intervals, $H = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$ where:

   $$h_{a,b,s} = \begin{cases} s & \text{if } x \in [a, b] \\ -s & \text{if } x \notin [a, b] \end{cases}$$

   b. The class of origin-centered spheres in R^d, $H = \{h_{a,s} : s \in \{-1, 1\}, a \in R\}$ where:

   $$h_{a,s} = \begin{cases} s & \text{if } x \text{ is within the origin centered sphere of radius } a \\ -s & \text{if } x \text{ is outside the origin centered sphere of radius } a \end{cases}$$

2. Consider two hypothesis spaces H_1, H_2 such that $H_1 \subset H_2$. Prove that the VC Dimension of H_2 is at least as large as the VC Dimension of H_1.

Solution

1. a. The VC Dimension is 3.
   Let the hypothesis space of intervals be $\mathcal{H}_i$. For any set of three points, all labellings (s, s, s), $(s, -s, s)$, $(-s, s, s)$, $(-s, -s, s)$ ... can be realized by hypothesis space $\mathcal{H}_i$ by placing $[a, b]$ to include all continuous points with the same label. Thus $VC(\mathcal{H}_i) \geq 3$. Now consider a set of four points. $\mathcal{H}_i$ cannot realize the labelling $(s, -s, s, -s)$, etc so $VC(\mathcal{H}_i) < 4$. Combining the two inequalities, we have that $VC(\mathcal{H}_i)$ is equal to 3, because the VC dimension must be an integer.

   b. The VC Dimension is 2.
   Let the hypothesis space of origin-centered spheres be $\mathcal{H}_s$. For any set of two points, we can either increase the radius or decrease it to make sure that both points are labelled correctly. Thus we know that $VC(\mathcal{H}_s) \geq 2$. Now consider a set of three points where the point closest to the origin has label s, the furthest point also has label s, and the point in between has label $-s$. In this case there is no such hypothesis $h \in \mathcal{H}_s$ that can realize this labelling. Thus $VC(\mathcal{H}_s) < 3$, and combining the inequalities, we have that $VC(\mathcal{H}_s) = 2$.

2. If you can shatter k points with one set of functions, adding more functions doesn't make it smaller. Certainly you should have proved it more rigorously but this is the general intuition that we were looking for.

Problem 4

(12 points)
In the following question, we will bound the VC dimension for a binary K-means classifier on a 1-dimensional space.

a. Let K be the space of K-means classifiers with k positive means and k negative means.

- Prove that some set C of size $2k$ can be shattered by K.

- Argue that there is no set of points C of size $2k+1$ that may be shattered by K

b. Is a polynomial (in k, d, $1/\epsilon$, and $1/\delta$) amount of data enough to determine a good K means classifier? Why or why not?

Solution

a
- **Proof 1** Consider a set of $2k$ equally spaced points on the number line. We would like to show that we can perfectly classify these $2k$ for all possible labelings.

  Starting at the smallest point, we can group adjacent points into k pairs. There are four possible labellings for each pair: (+ +), (+ -), (- +) , (- -). If a pair is labeled (+, +), place a positive center in the middle of the two points. Likewise, if a pair is labeled (-, -), place a negative center in the middle of the two points. If a pair is labeled either (+, -) or (-, +), place a positive center on the positive point and a negative center on the negative point. Notice that if points are each separated by some distance ϵ, that we are placing a center at most $\epsilon/2$ distance away. The closest one pair can be to another is ϵ and therefore, when we consider the pairs altogether no pair will "override" the classification of another pair (since it will not be closer to the point).

  Finally notice that each pair uses at most one positive center and/or at most one negative center. Since there are k pairs, then we use at most k positive centers and at most k negative centers.

  Proof 2 (By Induction) We will prove that there exists a set of size $2k$ that can be shattered with k positive centers and k negative centers by induction on k. As the base case, we can trivially see that we can shatter two points, with 1 positive center and 1 negative center. If the points have the same label, we put a center between them with that label. If they have different labels, we place a center at each point with the correct label. Notice that this uses at most 1 positive and at most 1 negative center.

  Now assume that there exists a set C_{k-1} of size $2(k-1)$ that we can shatter with $k-1$ positive centers and $k-1$ negative centers. For all labellings, calls the maximum distance between a center and the point it labels ϵ. Let C_k be the set C_{k-1} plus two new points added at "the nd" (after the furthest point in C_{k-1}) at least ϵ away from any of the points in C_{k-1}.

  Now, we would like to show that we can shatter the set C_k with k positive centers and k negative centers. Since we added the two new points at least ϵ away from the points in C_{k-1}, we will still be able to shatter C_{k-1} with $k-1$ positive centers and $k-1$ negative centers. We can shatter the two new points, the same way that we shattered the base case. Notice that this will only add at most one positive center and at most one negative center. Therefore, we have shown that there is a set (C_k) of size $2k$ that we can shatter with k positive centers and k negative centers.
- Consider any set of $2k+1$ points. If any two points are equal, then we can label them differently and they cannot be shattered. Therefore, let's consider $2k+1$ distinct points. Then, assign alternating labels to the $2k+1$ points (ex: $+, -, \cdots +$). Notice that there is no way to place the k positive centers and k negative centers so that we correctly classify all the data points.

b We can apply the bound from the VC-theorem (slide 31 of the continuous inputs section). We know that the VC dimension here is $2k = d$, so substituting we have

$$m \geq C\frac{d + \log(1/\delta)}{\epsilon^2} = C\frac{2k + \log(1/\delta)}{\epsilon^2}$$

From this, we see that the number of samples required is indeed polynomial in $k, 1/\delta$ and $1/\epsilon$.