SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

51.504 Machine Learning, Fall 2018

Assignment 2

Last update: Tuesday 9th October, 2018 04:24

Grading Policy and Due Date

- You are required to submit a report that summarizes your results and findings, based on each of the following question asked
- Submit your assignment report to eDimension.
- This assignment is an individual assignment. Discussions amongst yourselves are allowed and encouraged, but you should write your own report. Write down the names who you have talked to when doing the homework.
- Submit your assignment to eDimension by 14 October 2018 11.59pm. This is a hard deadline. Late submissions will be heavily penalized (20% deduction per day).

Task: K-means

In this problem we will look at the K-means clustering algorithm. Let $X = \{x_1, x_2, ..., x_n\}$ be our data and γ be an indicator matrix such that $\gamma_{ij} = 1$ if x_i belongs to the jth cluster and 0 otherwise. Let $\{\mu_1, \mu_2, ..., \mu_k\}$ be the means of the clusters. We can define the loss function L as follows,

$$L(\gamma, \mu_1, \mu_2, ..., \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} \|x_i - \mu_j\|^2. \tag{1}$$

We define $C = 1, 2, ..., k$ be the set of clusters. The most common form of the K-means algorithm proceeds as follows.

- Initialize $\mu_1, \mu_2, ..., \mu_k$.
- While L is decreasing, iterate the following.
  - Determine γ, breaking ties arbitrarily

$$\begin{cases} 1, & \|x_i - \mu_j\| \leq \|x_i - \mu_{j'}\|^2, \forall j' \\ 0, & otherwise \end{cases} \tag{2}$$

  - Recompute μ_j using the updated γ. Remove j from C if $\sum_{i=1}^{n} \gamma_{ij} = 0$. Otherwise,

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{\sum_{i=1}^{n} \gamma_{ij}} \tag{3}$$

1. (1 pt) Show that this algorithm will always terminate in a finite number of steps. (Hint: how many different values can γ take?)

2. (2 pt) Let $\hat{x}$ be the sample mean. Consider the following quantities,

$$T(X) = \frac{\sum_{i=1}^{n} \|x_i - \hat{x}\|^2}{n} \tag{4}$$

$$W_j(x) = \frac{\sum_{i=1}^{n} \gamma_{ij} \|x_i - \mu_j\|^2}{\sum_{i=1}^{n} \gamma_{ij}} \tag{5}$$

$$B(x) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} \|\mu_j - \hat{x}\|^2 \tag{6}$$

   Here, $T(x)$ is the total deviation, $W_j(X)$ is the intra-cluster deviation and $B(X)$ is the inter-cluster deviation. What is the relation between these quantities? Based on this, show that K-means can be views as minimizing a weighted average of intra-cluster deviation while approximately maximizing the inter-cluster deviation. Your relation may contain a term that was not mentioned above.

3. (1 pt) Show that the minimum of L is a non increasing function of k, the number of clusters. Argue that this means it is meaningless to choose the number of clusters by minimizing L.

Task: Support Vector Machine

1. (2 pt) Make an illustration with the following points: $\{((1,1),-1), ((1,3),-1), ((2,5),-1), ((4,2),-1), ((5,5),1), ((5,9),1), ((6,1),-1), ((7,5),1), ((9,1),1), ((11,3),1), ((11,7),1)\}$. Try to manually guess what should be the support vectors. Draw the support vector lines and the separating hyperplane (by hand, approximate lines).

2. (1 pt) Find the weight vector w and bias b of the separating hyperplane and represent it in the form $w_1x_1 + w_2x_2 + b = 0$.

3. (1 pt) We represent $f(x) = wx + b$. At the moment the value of function f in the support vectors are -1.5 and 1.5. This is completely OK, but in order to make future explanations more elegant and to make the optimization task for finding the suitable weight vector easier to write down, we can scale the weight vector and bias so that the support vectors would have function value of 1. Find the scaled w and b. You can do this by dividing the line formula by 1.5. Why?

4. (1 pt) There are two types of margins we can talk about. Functional margin and geometrical margin. Functional margin shows the function value of some point. Geometrical margin shows the actual distance of the point to the separating hyperplane. When we do scaling of the weights and bias as in task 3 then the functional margin will change (because the function changes), but geometrical margin stays the same (because the geometrical position of the separating hyperplane does not change with the scaling). Definitions of the two margins are:

   Functional margin: $y_i f(x_i)$

   Geometrical margin: $\frac{y_i f(x_i)}{\|w\|}$, this comes directly from the formula for finding point distance from line, except that the upper part is multiplied with the true label to achieve a sign for the margin indicating correctness of classification.

Find the functional and geometrical margin for the 4th point. We know that the geometrical margin shows the actual distance from point to line. But what does the functional margin show? What can you say about when the functional margin is positive and when it is negative?