CS536 Homework1

February 9, 2018

1 Problem

Consider a Gaussian in d dimensions with zero mean and spherical covariance $\sigma^2\mathbf{I}$.

1. Show that the integral of the probability density over a thin shell of radius r and thickness $\epsilon << 1$ is

$$p(r|d)\epsilon = \frac{S_d r^{d-1}}{(2\pi\sigma^2)^{d/2}} e^{-\frac{r^2}{2\sigma^2}}\epsilon,$$

   where S_d is the surface area of the unit sphere in d dimensions.

2. Show that the function $p(r|d)$ has a stationary point located at $r^* \approx \sqrt{d}\sigma$ for large d.

3. Plot $p(r^*|d)$ for $d = 1, 2, 5, 10, 20$.

4. What can you say about the relationship between the density at the origin and the maximum probability mass in a thin shell, as a function of the dimension d?

2 Problem

Consider the family of concepts in a 2D Euclidean plane $X = \Re^2$ consisting of concentric circles, $c = \{(x, y) : x^2 + y^2 \leq r^2\}$ for some $r \in \Re$. Show that this class can be (ϵ, δ)-PAC-learned from training data of size $N \geq (1/\epsilon)\log(1/\delta)$.

3 Problem

Consider the generalization error of the L_q loss

$$E[L_q(t, y(\boldsymbol{x}))] = \int_{t,y} |t - y(\boldsymbol{x})|^q p(t, \boldsymbol{x}) dt d\boldsymbol{x}.$$

Prove that the optimal estimator of $\boldsymbol{y}$ that minimizes this error is:

1. For $q = 2$:

$$y^*(\boldsymbol{x}) = \mathbb{E}_t[t|\boldsymbol{x}].$$

2. For $q = 1$:

$$y^*(\boldsymbol{x})) = \text{median}(t|\boldsymbol{x}),$$

   i.e., the function such that the probability mass of $t > y(\boldsymbol{x})$ is the same as the probability mass for $t \leq \boldsymbol{x}$.

3. For $q \to 0$:

$$y^*(\boldsymbol{x}) = \text{mode}(t|\boldsymbol{x}),$$

   i.e., the function of $y(\boldsymbol{x})$ that maximizes $p(t|\boldsymbol{x})$ over t for each $\boldsymbol{x}$.

4 Problem

In this problem we are going to implement the LDA for 20 Newsgroups data set. We are going to:

1. We begin by getting familiar with the data set and split the data set into two groups, i.e. training data and test data.
2. Then we use the one dimension reduction technique (i.e. PCA) to reduce the feature dimension since working with original data set will probably result in memory crash.
3. After reducing the feature dimension, we start building LDA model using the train data. In LDA we basically compute the model parameters by using formulas in (4:36), (4:37) and (4:38) in K. P. Murphy book with some minor changes to make the computations more efficient.
4. Finally, we predict the LDA with the computed parameters on the test data and compute its accuracy rate and compare its performance to other linear models built in SciKit-Learn package.

Please open the jupyter notebook called LDA_student.ipynb and start working on LDA problem by following the instruction.

5 Problem

5.1 KNN and 4 News group Classification

You will implement k-Nearest Neighbor (KNN) classifiers on the NEWS_DATASET . You are going to use the 4 classes ('alt.atheism','talk.religion.misc','comp.graphics','sci.space') among the original 20 classes.

5.1.1 Run: hw1_student → python features.py

5.1.2 Implement KNN algorithm and choose the parameter K.

1. Your working files are :
   - hw1_student → knn.ipynb
   - hw1_student → cs536_1 → models → k_nearest_neighbor.py
2. Fill out the working files with your code as appropriate. For the metric, please use the L2 norm.
3. When you finish your implementation, you will find validation accuracy rates for the different Ks at the **train_ratio:1.0** (controlling number of training dataset).
4. Now, you will change the **train_ratio** from 0.1 to 1.0 (interval step : 0.1) and guess the appropriate Ks for each value of the **train_ratio**s.
5. How did you choose the K? Explain your rule for the selection of the K.

5.1.3 Prediction on Test Set

1. Your working files :
   - hw1_student → knn_test.ipynb and
   - hw1_student → cs536_1 → models → k_nearest_neighbor.py
2. Fill out the working files with your code as appropriate.
3. When you finish your implementation, you will find test accuracy rates for the different Ks at the **train_ratio:1.0**.
4. Now, you will change the **train_ratio** from 0.1 to 1.0 (interval step : 0.1) and find the test_ratios.
5. Based on the your test accuracies, are your previous selections (validation set) still valid? If not, how would you change the selection of the parameter K for each **train_ratio**.

Table 1: Validation Accuracy

Train_Ratio	K								
	1	**3**	**5**	**7**	**9**	**11**	**15**	**19**	K **Selection**
0.1									
0.2									
0.3									
0.4									
0.5									
0.6									
0.7									
0.8									
0.9									
1.0									

Table 2: Test Accuracy

Train_Ratio	K							
	1	**3**	**5**	**7**	**9**	**11**	**15**	**19**
0.1								
0.2								
0.3								
0.4								
0.5								
0.6								
0.7								
0.8								
0.9								
1.0								

5.1.4 Analysis

Based on overall experiments results, please characterize the KNN along with the number of training data. Do you have any suggestion for better K selection in section 5.1.2 experiment.