CS 7641 CSE/ISYE 6740 Mid-term Exam Solution (2013 Fall)

Le Song

10/17 Thr, 1:35 - 2:55 pm

- Name:
- GT ID:
- E-mail:

Problem	Point	Your Score
1	15	
2	15	
3	15	
4	15	
5	20	
6	20	
Total	100	

Instructions:

- Try your best to be clear as much as possible. No credit may be given to unreadable writing.
- The exam is open book and open note, but no electronic devices (including smart phones) are allowed.
- Good luck!

1 Maximum Likelihood [15 pts]

You are playing a game with two coins. Coin 1 has a θ probability of heads. Coin 2 has a 2θ probability of heads. You flip these coins several times and record your results:

Coin	Result
1	Head
2	Tail
2	Tail
2	Tail
2	Head

(a) What is the likelihood of the data given θ? [6 pts]

Answer:

$$\begin{aligned} L(\theta) &= p(\text{data}|\theta) \\ &= p(\text{Coin1} = \text{Head})[p(\text{Coin2} = \text{Tail})]^3 p(\text{Coin2} = \text{Head}) \\ &= \theta(1-2\theta)^3 2\theta \\ &= 2\theta^2(1-2\theta)^3 \end{aligned}$$

(b) What is the maximum likelihood estimation for θ? [4 pts]

Answer:

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \log 2 + 2\log\theta + 3\log(1-2\theta)^3 \\ \frac{\partial \ell(\theta)}{\partial \theta} &= \frac{2}{\theta} + \frac{3\cdot -2}{1-2\theta} \\ &= 2(1-2\theta) - 6\theta = 0 \\ \Rightarrow \quad \theta_{MLE} &= \frac{1}{5} \end{aligned}$$

(c) Uniform distribution [5 pts]

A uniform distribution in the range of $[0, \theta]$ is given by

$$p(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}.$$

What is the maximum likelihood estimator of θ?

Hint: think of two cases, where $\theta < \max(x^i)$ and $\theta \ge \max(x^i)$ separately.

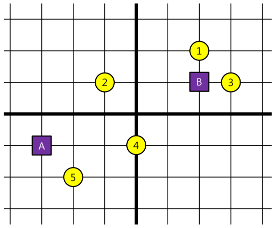
Answer:

Suppose $\theta < \max(x^i)$. If the data is uniformly distributed between $[0, \theta]$, it is impossible to get any data point larger than θ. Thus, this case is impossible.

Now suppose $\theta \ge \max(x^i)$. As the distribution is uniform, larger θ would lower the likelihood of observed data points. Thus, maximum likelihood of θ arises with the tightest setting, $\theta_{MLE} = \max(x^i)$.

2 K-means [15 pts]

The following figure shows an intermediate state of running K-means. Circles are data points, and squares are centroid. We assume two-dimensional Euclidean space, where the point that two thick lines are crossing is the origin. For example, the centroid A is in (-3, 1), and the data point 2 is in (-1, 1).



(a) Suppose we use Euclidean distance $\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$ for the distance measure between two points (x_1, x_2) and (y_1, y_2). Which data points are belonging to cluster A and B, respectively? [4 pts]

- Cluster A: 2, 5
- Cluster B: 1, 3, 4

(b) Suppose we proceed centroid recalculation step from the above figure. Where are those two centroids located after this step? [4 pts]

- Centroid A: $(-\frac{3}{2}, -\frac{1}{2})$
- Centroid B: $(\frac{5}{3}, \frac{2}{3})$

(c) After the step in (b), is the K-means iteration done? Mark one of the following. [3 pts]

- Yes / **No**

(d) Suppose we use Manhattan distance $|x_1-y_1|+|x_2-y_2|$ for the distance measure between two points (x_1, x_2) and (y_1, y_2). Which data points are belonging to cluster A and B, respectively? Is this same with (a)? [4 pts]

- Cluster A: 4, 5
- Cluster B: 1, 2, 3

No, it is not the same.

3 Principal Component Analysis [15 pts]

Suppose we have three data points in the two-dimensional Euclidean space, $(1, 1)$, $(2, 2)$ and $(3, 3)$.

(a) Calculate the first principle component. [5 pts]

Answer:

The first component is along the direction by moving the origin to the point $(2, 2)$. By normalizing, we have $(\sqrt{2}/2, \sqrt{2}/2)^\top$ (the negation is also correct).

(b) If we want to project the original data points into the 1-D space by the principle component you found in (a), what is the variance of the projected data? [5 pts]

Answer:

Along the direction $(\sqrt{2}/2, \sqrt{2}/2)^\top$, the projected coordinates are $(-\sqrt{2}, 0), (0, 0), (\sqrt{2}, 0)$, so the variance will be $4/3$.

(c) What is the reconstruction error when we reduce the dimension of given data from 2-D to 1-D? [5 pts]

Answer:

The reconstruction error is 0, as all data points are on the principal component. We lose no information (variance).

4 Generative and Discriminative Classifiers [15 pts]

(a) Suppose we have a customer who needs our help to classify job applications into good/bad categories, and at the same time to detect job applicants who lie in their applications using density estimation to detect outliers. To meet her needs, do you recommend using a discriminative or generative classifier? Why? [5 pts]

Answer:
We should use generative classifier for that it gives us the probability density so as to find outliers later.

(b) In class, we have learned that *maximum likelihood (MLE)* learns the parameters by maximizing the joint likelihood of the given data. Alternatively, the *maximum a posteriori (MAP)* approach obtains a point estimate of the parameter by maximizing the posteriori likelihood, which incorporates a prior distribution over the parameter one wants to learn with the likelihood of the given data based on Bayes rule.

- If we try to predict using a Naive Bayes classifier, which one should we use? [5 pts]

Answer:

Naive Bayes is a generative classifier. It selects the class which maximizes MAP using Bayes rule.

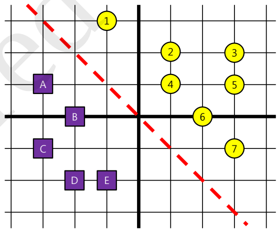
- If we try to learn the parameters of Logistic Regression, which principle to use? [5 pts]

Answer:

Logistic Regression maximizes MLE without referring to prior distribution.

5 Support Vector Machines [20 pts]

[a-c] The figure below shows a small dataset with two classes (circles and rectangles). The red dotted line indicates the decision boundary we found using a support vector machine. We assume two-dimensional Euclidean space, where the point that two thick lines are crossing is the origin. For example, A is in (-3, 1), and the data point 6 is in (2, 0).



(a) List all support vectors. [4 pts]

- Answer: 1, 4, 6, 7, A, B

(b) Suppose we have an additional square class data point F at (1, 0). When we learn an SVM including this data point, will the decision boundary be the same? [4 pts]

- Yes / **<u>No</u>**

(c) Suppose we have an additional circle class data point 8 at (1, 0). When we learn an SVM including this data point, will the decision boundary be the same? [4 pts]

- Yes / **<u>No</u>**

[d-e] suppose we have a dataset with 25 training points and 10 test points for a binary classification problem. Among the 25 training points, 15 of them are for class A and 10 are for class B. Also, among the 10 test points, 5 of them are for class A and the rest are for class B.

(d) What is the maximum number of possible support vectors with this dataset? [4 pts]

- Answer: 25

(e) What is the minimum number of possible support vectors with this dataset? [4 pts]

- Answer: 2

6 True/False [20 pts]

Please mark on T if the statement is true, or F if it is not always true. No need to explain why. [2 pts for correct answer, 0 pts for no answer, -2 pts for wrong answer, each]

(a) Principal component analysis preserves variance as much as possible.

Answer: T

(b) Clustering using K-means algorithm is a supervised learning task.

Answer: F

(c) Non-parametric density models do not have parameters.

Answer: F

(d) Naive Bayes does not work well if its independence assumption is not satisfied.

Answer: F

(e) Bayes decision rule is the theoretically best classifier that minimize probability of classification error.

Answer: T

(f) Logistic regression is a generative classifier.

Answer: F

(g) K-means usually converges to a local optimum, but EM algorithm guarantees convergence to the global optimum.

Answer: F

(h) With dual representation of SVM, we need only the inner products (or kernel) between the examples.

Answer: T

(i) Bayesian treats a parameter as a fixed, unknown constant, not a random variable.

Answer: F

(j) A smoothing kernel is a multi-modal function.

Answer: F