CS 7641 CSE/ISYE 6740 2015 Final Exam Sample Question

Le Song

12/7 Mon

1 Maximum Likelihood [10 pts]

(a) Pareto Distribution [5 pts]

The Pareto distribution has been used in economics as a model for a density function with a slowly decaying tail:

$$f(x|x_0,\theta) = \theta x_0^{\theta} x^{-\theta-1},\ x \geq x_0,\ \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, X_2, \cdots, X_n$ is an i.i.d. sample. Find the maximum likelihood estimator of θ.

Answer: The log-likelihood function is

$$l(\theta) = n\log\theta + n\theta\log x_0 - (\theta+1)\sum_{i=1}^{n}\log X_i$$

Let the derivative with respect to θ be zero, we have

$$\hat{\theta} = \frac{1}{\overline{\log X} - \log x_0}$$

(b) Dependent Noise Model [5 pts]

Let X_1, X_2 be 2 determinations of a physical constant θ. Consider the model,

$$X_i = \theta + e_i, \quad i = 1, 2$$

and assume

$$e_i = \alpha e_{i-1} + \epsilon_i, \quad i = 1, 2, \quad e_0 = 0$$

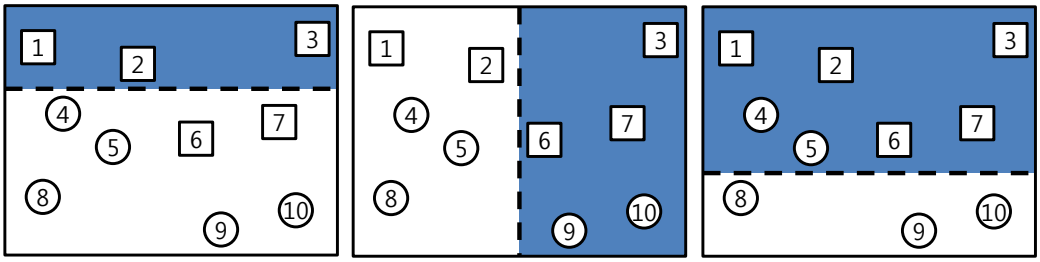
with ϵ_i is i.i.d standard normal, and α is a known constant. What is the maximum likelihood estimate of θ?

Answer:

$$\hat{\theta} = \frac{x_1 + \alpha(\alpha-1)x_1 - (\alpha-1)x_2}{(\alpha-1)^2 + 1}$$

2 Boosting [20 pts]

In this problem, we test your understanding of AdaBoost algorithm with a simple binary classification example. We are given 10 data points, belonging to either the `square` class or the `circle` class. The following figures show the decision boundary of three weak learners (h_1, h_2, and h_3). Suppose we boost with these weak learners in that order. In the figure, the darkened region means it is classified as `square` class, while white region indicates it is classified as `circle` class by the corresponding weak learner.



(a) When we learn the second weak learner h_2, list data points which receives higher weights than others. [4 pts]

Answer: 6, 7

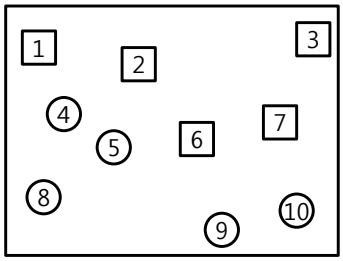
(b) When we learn the third weak learner h_3, list data points which receive smallest weights. [4 pts]

Answer: 3, 4, 5, 8

(c) What are ϵ_i for $i = 1, 2, 3$? [4 pts]

Answer: 0.2, 0.4, 0.2

(d) Draw the final decision boundary in the figure below. [4 pts]



(e) What is the classification error with train data of the final classifier? [4 pts]

Answer: 0

3 Model Selection [20 pts]

We learned about bias-variance decomposition and model selection in the class. It basically deals with the problem of choosing a model complexity. In each sub-questions, we show a pair of two candidate models for various machine learning problems. Mark B on the one with less bias, and mark V on the other. You are not required to explain why. For example,

- Model 1: A model with less bias – (B)
- Model 2: A model with less variance – (V)

(a) [3 pts]

- Model 1: A flexible model with many parameters – ( )
- Model 2: A rigid model with a few parameters – ( )

  Answer: B V

(b) [3 pts]

- Model 1: Ridge regression with large regularization coefficient λ – ( )
- Model 2: Unregularized linear regression – ( )

  Answer: V B

(c) [3 pts]

- Model 1: Polynomial regression model with higher degree – ( )
- Model 2: Polynomial regression model with lower degree – ( )

  Answer: B V

(d) [3 pts] For a logistic regression $p(y = 1|x, \theta) = \frac{1}{1+\exp\{-\theta^\top x\}}$**,**

- Model 1: Logistic regression with large θ – ( )
- Model 2: Logistic regression with small θ – ( )

  Answer: B V

(e) [4 pts]

- Model 1: K-nearest Neighbor with large K – ( )
- Model 2: K-nearest Neighbor with small K – ( )

  Answer: VB

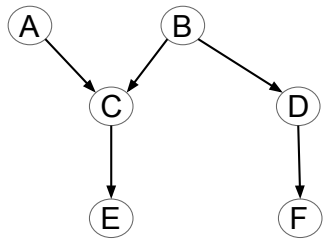
(f) [4 pts]

- Model 1: A classifier with irregular decision boundary – ( )
- Model 2: A classifier with smooth decision boundary – ( )

  Answer: BV

4 Graphical Models [10 pts]

Consider the following bayesian network with six variables -



(a) Factorize the joint probability distribution in terms of conditional probabilities based on chain rule. [3 pts]

Answer: $P(A)P(B)P(C|A,B)P(D|B)P(F|D)P(E|C)$

(b) Which of the following independence properties always hold in this model? [3 pts]

- $(A \perp B|C)$
- $(C \perp D|B)$
- $(E \perp A|C)$
- $(F \perp E)$

Answer: $(C \perp D|B)$ and $(E \perp A|C)$

(c) If we change the network to a markov network, will the factorization of the joint probability change? [2 pts]

Answer: Yes

(d) If we change the network to a markov network, will the independence properties change? [2 pts]

Answer: Yes

5 Support Vector Machine [20 pts]

Suppose we have a dataset in $1-d$ space which consists of 3 data points $\{-1, 0, 1\}$ with the positive label and 3 data points $\{-3, -2, 2\}$ with the negative label.

(a) Find a feature map ($\mathbb{R}^1 \to \mathbb{R}^2$), which will map the original $1-d$ data points to the $2-d$ feature space so that the positive samples and the negative samples are linearly separable with each other. Draw the dataset after mapping in the $2-d$ space. [10 pts]

(b) In your plot, draw the decision boundary given by hard-margin linear SVM. Mark the corresponding support vectors. [5 pts]

(c) For the feature map you use, what is the corresponding kernel $K(x_1, x_2)$? [5 pts]