Learning in Parametric Models

9/9 得分 (100%)

返回第 1 周课程

测验通过！

1 / 1 分

1. **Computing Sufficient Statistics.** Suppose that you are playing Dungeons & Dragons, and you suspect that the 4-sided die that your Dungeon Master is using is biased. In the past 60 times that you have attacked with your dagger and the 4-sided die was rolled to calculate how many hit points of damage you inflicted, 20 times it has come up 1, 15 times it has come up 2, 15 times it has come up 3, and 10 times it has come up 4.

Let θ_1 be the true probability of the die landing on 1 , and similarly for θ_2 , θ_3 , and θ_4 . You want to estimate these parameters from the past 60 rolls that you observed using a simple multinomial model. Which of the following is a sufficient statistic for this data?

- [x] A vector with with four components, with the i^{th} component being the number of times you dealt i hit points worth of damage.

正确
A sufficient statistic is a function of the data that summarizes the relevant information for computing the likelihood. The sufficient statistics for a multinomial model are the "counts" of each possible result. The number of times each digit was rolled allows us to compute the likelihood function.

- [ ] The total number of times you viciously attacked a monster with your dagger (i.e., the total number of times that the dice was rolled).
- [ ] None of these are sufficient statistics.
- [ ] The total amount of damage you inflicted with your trusty dagger (i.e., the sum of all die rolls).

1 / 1 分

2. **MLE Parameter Estimation.** In the context of the previous question, what is the unique Maximum Likelihood Estimate (MLE) of the parameter θ_1 ? Give your answers rounded to the nearest ten-thousandth (i.e. 1/6 should be 0.1667).

0.3333

正确回答

1 / 1 分

3. **Likelihood Functions.** For a Naive Bayes network with one parent node, X , and 3 children nodes, Y_1, Y_2, Y_3 , which of the expressions below would be a correct expression for the likelihood, decomposed in terms of the local likelihood functions?

- [ ] $L(\theta : D) = \prod_{m=1}^{M} P(x[m], y_1[m], y_2[m], y_3[m] : \theta)$
- [ ] $L(\theta : D) = \prod_{m=1}^{M} P(y_1[m]|x[m] : \theta_{Y_1|X}) P(y_2[m]|x[m] : \theta_{Y_2|X}) P(y_3[m] : \theta_{Y_2|X}) P(x[m]|y_1[m], y_2[m], y_3[$
- [ ] $L(\theta : D) = \prod_{m=1}^{M} P(y_1[m], y_2[m], y_3[m] : \theta) P(x[m]|y_1[m], y_2[m], y_3[m] : \theta)$
- [x] $L(\theta : D) = (\prod_{m=1}^{M} P(x[m] : \theta_X))(\prod_{m=1}^{M} P(y_1[m]|x[m] : \theta_{Y_1|X}))(\prod_{m=1}^{M} P(y_2[m]|x[m] : \theta_{Y_2|X}))(\prod_{m=1}^{M} P$

正确
The formulation for a likelihood function decomposed into local likelihood functions follows directly from the lecture videos and takes this form.

1 / 1 分

4. **MLE for Naive Bayes.** Using a Naive Bayes model for spam classification with the vocabulary V={"SECRET", "OFFER", "LOW", "PRICE", "VALUED", "CUSTOMER", "TODAY", "DOLLAR", "MILLION", "SPORTS", "IS", "FOR", "PLAY", "HEALTHY", "PIZZA"}. We have the following example spam messages SPAM = {"MILLION DOLLAR OFFER", "SECRET OFFER TODAY", "SECRET IS SECRET"} and normal messages, NON-SPAM = {"LOW PRICE FOR VALUED CUSTOMER", "PLAY SECRET SPORTS TODAY", "SPORTS IS HEALTHY", "LOW PRICE PIZZA"}.

We create a multinomial naive Bayes model for the data given above. This can be modeled as a parent node taking values SPAM and NON-SPAM and a child node for each word in the vocabulary. The θ values are estimated based on the number of times that a word appears in the vocabulary. Give the MLE for θ_{SPAM} . Enter the value as a decimal rounded to the nearest ten-thousandth (0.xxxx).

0.4286

正确回答

1 / 1 分

5. **MLE for Naive Bayes.** Using the same data and model above, give the MLE for $\theta_{\text{SECRET}|\text{SPAM}}$. Enter the value as a decimal rounded to the nearest ten-thousandth (0.xxxx).

0.3333

正确回答

1 / 1 分

6. **MLE for Naive Bayes.** Using the same data and model above, give the MLE for $\theta_{\text{SECRET}|\text{NON-SPAM}}$. Enter the value as a decimal rounded to the nearest ten-thousandth (0.xxxx).

0.0667

正确回答

1 / 1 分

7. **Learning Setups.** Consider the following scenario: You have been given a dataset that contains patients and their gene expression data for 10 genes. You are also given a 0/1 label where 1 means that patient has disease A and 0 means the patient does not.

Your goal is to learn a classification algorithm that could predict these labels with high accuracy. You split the data into three sets:

1: Set of patients used for fitting the classifier parameters (e.g., the weights and bias of a logistic regression classifier).

2: Set of patients used for tuning the hyperparameters of the classifier (e.g., how much regularization to apply).

3: Set of patients used to assess the performance of the classifier.

What are these sets called?

- ○ 1 & 2: Training Set, 3: Validation Set.
- ◉ 1: Training Set, 2: Validation Set, 3: Test Set

正确

We fit parameters on training set, tune on validation set and assess performance on test set.

- 1: Validation Set, 2: Test Set, 3: Training Set.
- 1: Training Set, 2: Test Set, 3: Validation Set

✔ 1 / 1 分

8. **Constructing CPDs.** Assume that we are trying to construct a CPD for a random variable whose value labels a document (e.g., an email) as belonging to one of two categories (e.g., spam or non-spam). We have identified K words whose presence (or absence) in the document each changes the distribution over labels (e.g., the presence of the word "free" is more likely to indicate that the email is spam). Assume that we have M labeled documents that we use to estimate the parameters for the CPD of the label given indicator variables representing the appearance of words in the document. We plan to use maximum likelihood estimation to select the parameters of this CPD.

If $M = 1,000$ and $K = 30$, which of the following CPD types are most likely to provide the best generalization performance to unseen data? Mark all that apply.

- [ ] A linear Gaussian CPD

未选择的是正确的

- [ ] A table CPD

未选择的是正确的

- [x] A sigmoid CPD

正确
With a sigmoid CPD the number of parameters that will need to be learned is $K = 30$ (plus 1 for the bias term) and thus $M = 1000$ instances are sufficient to get a reasonable maximum likelihood estimation of the parameters and hence the distribution.

- [ ] None of these CPDs would work.

未选择的是正确的

✔ 1 / 1 分

9. **Constructing CPDs.** For the same scenario as described in the previous question,

if $M = 100,000$ and $K = 3$, which of the following CPD types is most likely to provide the best generalization performance to unseen data?

- A tree CPD with $K = 3$ leaves
- A sigmoid CPD
- A linear Gaussian CPD
- (●) A table CPD

正确
In this scenario, a table CPD has $(2^3 - 1)$ free parameters, so we have enough instances to get a good estimate of the distribution for this type of CPD.
