



# Probabilistic Graphical Models

Daphne Koller, Kevin Murphy  
Winter 2011-2012

[Home](#)
[Quizzes](#)
[Theory Problems](#)
[Assignments](#)
[Assignment Questions](#)
[Video Lectures](#)
[Discussion Forums](#)
[Course Wiki](#)
[Lecture Slides](#)
[Course Schedule](#)
[Course Logistics](#)
[Course Information](#)
[Course Staff](#)
[Octave Installation](#)


## Feedback — Learning in Parametric Models

You achieved a score of 7.00 out of 7.00

### Question 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  $\theta_1$  be the true probability of the die landing on 1, and similarly for  $\theta_2$ ,  $\theta_3$ , and  $\theta_4$ .

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?

Your Answer	Score	Explanation
<input checked="" type="radio"/> A vector with with four components, with the $i^{th}$ component being the number of times you dealt $i$ hit points worth of damage.	 1.00	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.
Total	1.00 / 1.00	

## Question 2





**MLE Parameter Estimation.** In the context of the previous question, what is the unique Maximum Likelihood Estimate (MLE) of the parameters  $\theta$ ? Enter  $\theta_1$  on the first line,  $\theta_2$  on the second, and so forth (til  $\theta_4$ ). Give your answers rounded to the nearest ten-thousandth (i.e.  $1/3$  should be 0.3333).

0.3333

0.25





Your Answer		Score	Explanation
0.3333		0.25	
0.25		0.25	
0.25		0.25	
0.1666		0.25	
Total		1.00 / 1.00	

## Question 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?

Your Answer

☒

$$L(\theta : D) = (\prod_{m=1}^M P(x[m] : \theta_X)) (\prod_{m=1}^M P(y_1[m] | x[m] : \theta_{Y_1|X})) (\prod_{m=1}^M P(y_2[m] | x[m] : \theta_{Y_2|X})) (\prod_{m=1}^M P(y_3[m] | x[m] : \theta_{Y_3|X}))$$

---

Total

---

#### Question 4

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

We create 2 separate Naive-Bayes generative models, one for SPAM and another for NON-SPAM using the data given above. This can be modeled as a parent node taking values SPAM and NON-SPAM and a child node that represents the presence of a single word from the vocabulary in the message. Give the MLEs for  $\theta_{\text{SPAM}}$ ,  $\theta_{\text{SECRET}|\text{SPAM}}$ ,  $\theta_{\text{SECRET}|\text{NON-SPAM}}$ ,  $\theta_{\text{SPORTS}|\text{NON-SPAM}}$ ,  $\theta_{\text{DOLLAR}|\text{SPAM}}$  respectively. Separate each with new lines, in the order listed above. Enter the value as a decimal rounded to the nearest ten-thousandth (0.xxxx).

0.4285  
0.3333




Your Answer	Score	Explanation
0.4285	0.20	Recall for naive bayes models, our theta parameter is just the proportion of samples with the specified characteristic out of all samples that match the given characteristics. I.e. $\theta_{A B}$ is the proportion of samples of type $A$ out of those matching description $B$ .
0.3333	0.20	Recall for naive bayes models, our theta parameter is just the proportion of samples with the specified characteristic out of all samples that match the given characteristics. I.e. $\theta_{A B}$ is the proportion of samples of type $A$ out of those matching description $B$ .
0.0666	0.20	Recall for naive bayes models, our theta parameter is just the proportion of samples with the specified characteristic out of all samples that match the given characteristics. I.e. $\theta_{A B}$ is the proportion of samples of type $A$ out of those matching description $B$ .
0.1333	0.20	Recall for naive bayes models, our theta parameter is just the proportion of samples with the specified characteristic out of all samples that match the given characteristics. I.e. $\theta_{A B}$ is the proportion of samples of type $A$ out of those matching description $B$ .
0.1111	0.20	Recall for naive bayes models, our theta parameter is just the proportion of samples with the specified characteristic out of all samples that match the given characteristics. I.e. $\theta_{A B}$ is the proportion of samples of type $A$ out of those matching description $B$ .
Total	1.00 / 1.00	

## Question 5

**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?





Your Answer	Score	Explanation
<input checked="" type="radio"/> 1: Training Set, 2: Validation Set, 3: Test Set	 1.00	We fit parameters on training set, tune on validation set and assess performance on test set.
Total	1.00 / 1.00	

## Question 6

**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.

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> A sigmoid CPD	 0.25	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.
<input type="checkbox"/> None of these CPDs would work.	 0.25	One of these CPDs would likely provide better generalization performance than the others.
<input type="checkbox"/> A linear Gaussian CPD	 0.25	A linear Gaussian CPD is inappropriate because the label variable is discrete (and in fact, binary) as it would take on the values "present" and "not present".
<input type="checkbox"/> A table CPD	 0.25	A table CPD has $(2^{30} - 1)$ free parameters and hence we do not have enough instances to get a reasonable estimate of the distribution for this type of CPD (the variance of our estimator will be high).
Total	1.00 / 1.00	

### Question 7

**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?

Your Answer	Score	Explanation
<input checked="" type="radio"/> A table CPD	 1.00	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.
Total	1.00 / 1.00	