EXERCISES 21

LEARNING PROBABILISTIC MODELS

21.1 Statistical Learning

Exercise 21.BAYC

The data used for Figure 21.1 on page 774 can be viewed as being generated by h_5. For each of the other four hypotheses, generate a data set of length 100 and plot the corresponding graphs for $P(h_i \mid d_1, \ldots, d_N)$ and $P(D_{N+1} = lime \mid d_1, \ldots, d_N)$. Comment on your results.

The code for this exercise is a straightforward implementation of Equations 21.1 and 21.2. Figure S21.1 shows the results for data sequences generated from h_3 and h_4. (Plots for h_1 and h_2 are essentially identical to those for h_5 and h_4.) Results obtained by students may vary because the data sequences are generated randomly from the specified candy distribution. In (a), the samples very closely reflect the true probabilities and the hypotheses other than h_3 are effectively ruled out very quickly. In (c), the early sample proportions are somewhere between 50/50 and 25/75; furthermore, h_3 has a higher prior than h_4. As a result, h_3 and h_4 vie for supremacy. Between 50 and 60 samples, a preponderance of limes ensures the defeat of h_3 and the prediction quickly converges to 0.75.

Exercise 21.BAYD

Repeat Exercise BAYES-CANDY-EXERCISE, this time plotting the values of $P(D_{N+1} = lime \mid h_{\text{MAP}})$ and $P(D_{N+1} = lime \mid h_{\text{ML}})$.

(Plots not shown.) plots are shown in Figure S**??**. Because both MAP and ML choose exactly one hypothesis for predictions, the prediction probabilities are all 0.0, 0.25, 0.5, 0.75, or 1.0. For small data sets the ML prediction in particular shows very large variance.

Exercise 21.BAYE

Suppose that Ann's utilities for cherry and lime candies are c_A and ℓ_A, whereas Bob's utilities are c_B and ℓ_B. (But once Ann has unwrapped a piece of candy, Bob won't buy it.) Presumably, if Bob likes lime candies much more than Ann, it would be wise for Ann to sell her bag of candies once she is sufficiently sure of its lime content. On the other hand, if Ann unwraps too many candies in the process, the bag will be worth less. Discuss the problem of determining the optimal point at which to sell the bag. Determine the expected

utility of the optimal procedure, given the prior distribution from Section 21.1.

This is a nontrivial sequential decision problem, but can be solved using the tools devel-oped in the book. It leads into general issues of statistical decision theory, stopping rules, etc. Here, we sketch the "straightforward" solution.

We can think of this problem as a simplified form of POMDP (see Chapter 16). The "belief states" are defined by the numbers of cherry and lime candies observed so far in the sampling process. Let these be C and L, and let $U(C, L)$ be the utility of the corresponding belief state. In any given state, there are two possible decisions: $sell$ and $sample$. There is a simple Bellman equation relating Q and U for the sampling case:

$$Q(C, L, sample) = P(cherry|C, L)U(C + 1, L) + P(lime|C, L)U(C, L + 1)$$

Let the posterior probability of each h_i be $P(h_i|C, L)$, the size of the bag be N, and the fraction of cherries in a bag of type i be f_i. Then the value obtained by selling is given by

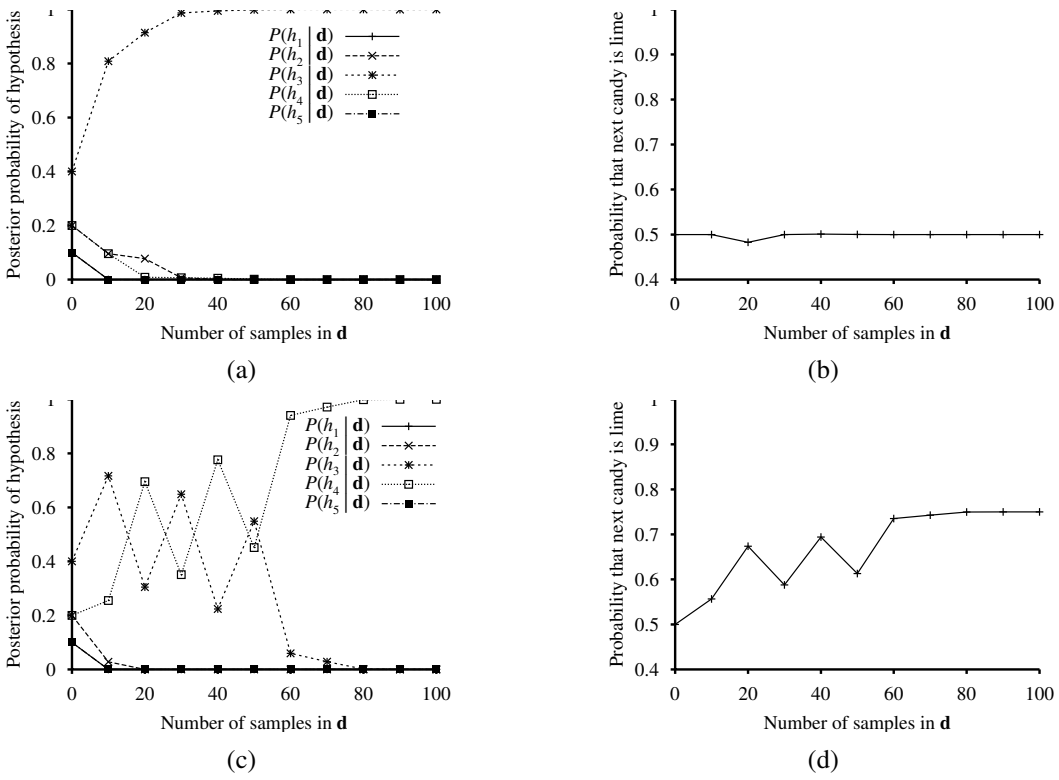


Figure S21.1 Graphs for Ex. 21.1. (a) Posterior probabilities $P(h_i|d_1, \ldots, d_N)$ over a sample sequence of length 100 generated from h_3 (50% cherry + 50% lime). (b) Bayesian prediction $P(d_{N+1} = lime|d_1, \ldots, d_N)$ given the data in (a). (c) Posterior probabilities $P(h_i|d_1, \ldots, d_N)$ over a sample sequence of length 100 generated from h_4 (25% cherry + 75% lime). (d) Bayesian prediction $P(d_{N+1} = lime|d_1, \ldots, d_N)$ given the data in (c).

the value of the sampled candies (which Ann gets to keep) plus the price paid by Bob (which equals the expected utility of the remaining candies for Bob):

$$Q(C, L, sell) = Cc_A + L\ell_A + \sum_i P(h_i|C, L)[(f_i N - C)c_B + ((1 - f_i)N - L)\ell_B]$$

and of course we have

$$U(C, L) = \max\{Q(C, L, sell), Q(C, L, sample)\} .$$

Thus we can set up a dynamic program to compute Q given the obvious boundary conditions for the case where $C + L = N$. The solution of this dynamic program gives the optimal policy for Ann. It will have the property that if she should sell at (C, L), then she should also sell at $(C, L + k)$ for all positive k. Thus, the problem is to determine, for each C, the threshold value of L at or above which she should sell. A minor complication is that the formula for $P(h_i|C, L)$ should take into account the non-replacement of candies and the finiteness of N, otherwise odd things will happen when $C + L$ is close to N.

Exercise 21.DRST

Two statisticians go to the doctor and are both given the same prognosis: A 40% chance that the problem is the deadly disease A, and a 60% chance of the fatal disease B. Fortunately, there are anti-A and anti-B drugs that are inexpensive, 100% effective, and free of side-effects. The statisticians have the choice of taking one drug, both, or neither. What will the first statistician (an avid Bayesian) do? How about the second statistician, who always uses the maximum likelihood hypothesis?

The doctor does some research and discovers that disease B actually comes in two versions, dextro-B and levo-B, which are equally likely and equally treatable by the anti-B drug. Now that there are three hypotheses, what will the two statisticians do?

The Bayesian approach would be to take both drugs. The maximum likelihood approach would be to take the anti-B drug. In the case where there are two versions of B, the Bayesian still recommends taking both drugs, while the maximum likelihood approach is now to take the anti-A drug, since it has a 40% chance of being correct, versus 30% for each of the B cases. This is of course a caricature, and you would be hard-pressed to find a doctor, even a rabid maximum-likelihood advocate who would prescribe like this. But you can find ones who do research like this.

21.2 Learning with Complete Data

Exercise 21.MISC.A

You have classification data with classes $Y \in \{+1, -1\}$ and features $F_i \in \{+1, -1\}$ for $i \in \{1, \ldots, K\}$. Say you duplicate each feature, so now each example has $2K$ features, with $F_{K+i} = F_i$ for $i \in \{1, \ldots, K\}$. Compare the *original* feature set with the *doubled* one and

indicate whether the below statements are true or false for Naïve Bayes. You may assume that in the case of ties, class $+1$ is always chosen. Assume that there are equal numbers of training examples in each class.

(i) The test accuracy could be higher with the original features.

(ii) The test accuracy could be higher with the doubled features.

(iii) The test accuracy will be the same with either feature set.

(iv) On a given training instance, the conditional probability $P(Y|F_1, \ldots)$ on a training instance could be more extreme (i.e. closer to 0 or 1) with the original features.

(v) On a given training instance, the conditional probability $P(Y|F_1, \ldots)$ on a training instance could be more extreme (i.e. closer to 0 or 1) with the doubled features.

(vi) On a given training instance, the conditional probability $P(Y|F_1, \ldots)$ on a training instance will be the same with either feature set.

(i) True.

(ii) False.

(iii) False.

(iv) False.

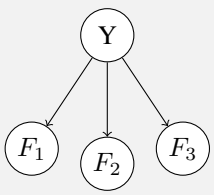
(v) True.

(vi) False.

Naïve Bayes makes the conditional independence assumption that all features are independent given the class label. Redundant features lead to "overconfidence" that may result in errors.

Exercise 21.MISC.F

Consider training the Naive Bayes model shown on the left with the training data provided in the table on the right.



F_1	0	0	1	0	1	1	1	1
F_2	0	1	0	1	1	0	1	1
F_3	1	1	1	0	0	1	1	0
Y	0	0	0	1	1	0	0	1

Calculate the maximum likelihood estimate of $P(F_1 = 1 \mid Y = 0)$.

$\frac{3}{5}$. This is found by counting the samples. There are 5 samples where $Y = 0$, and $F_1 = 1$ in 3 of them.

Exercise 21.XXXX

Your friend claims that he can write an effective Naive Bayes spam detector with only three features: the hour of the day that the email was received ($H \in \{1, 2, \ldots, 24\}$), whether it contains the word 'viagra' ($W \in \{\text{yes}, \text{no}\}$), and whether the email address of the sender is Known in his address book, Seen before in his inbox, or Unseen before ($E \in \{\text{K}, \text{S}, \text{U}\}$).

a. Flesh out the following information about this Bayes net:

(i) Graph structure.
(ii) Parameters.
(iii) Size of the set of parameters.

Suppose now that you labeled three of the emails in your mailbox to test this idea:

spam or ham?	H	W	E
spam	3	yes	S
ham	14	no	K
ham	15	no	K

b. Use the three instances to estimate the maximum likelihood estimate of the parameters.

c. Using the maximum likelihood parameters, find the predicted class of a new datapoint with $H = 3, W = \text{no}, E = U$.

d. You observe that you tend to receive spam emails in batches. In particular, if you receive one spam message, the next message is more likely to be a spam message as well. Explain a new graphical model which most naturally captures this phenomena.

(i) Graph structure.
(ii) Parameters.
(iii) Size of the set of parameters.

[]

a. (i) Graph structure: Naive Bayes net with three leaves
(ii) Parameters: $\theta_{spam}, \theta_{H,i,c}, i \in \{1, \ldots, 23\}, \theta_{W,c}, \theta_{E,j,c}, j \in \{K, S\}, c \in \{spam, ham\}$
(iii) Size of the set of parameters: $1 + 23 \cdot 2 + 2 + 2 \cdot 2$.

b.
- $\theta_{spam} = \frac{1}{3}$
- $\theta_{H,3,spam} = 1$
- $\theta_{H,14,ham} = \frac{1}{2}$
- $\theta_{H,15,ham} = \frac{1}{2}$
- $\theta_{W,spam} = 1$
- $\theta_{E,S,spam} = 1$
- $\theta_{E,K,ham} = 1$

c. Both assign a likelihood of zero, so no prediction is specified by the maximum likelihood parameters.

d. (i) Graph structure: A HMM, except that there are three observations at each node.
(ii) Parameters: Add 2 parameters: transition to spam from spam and from ham.

(iii) Size of the set of parameters: Add 2 to the expression in the first question.

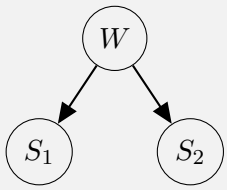
Exercise 21.XXXX

Stoplights S_1 and S_2 can each be in one of two states: green (g) or red (r). Additionally, the machinery behind both stoplights (W) can be in one of two states: working (w) or broken (b). We collect data by observing the stoplights and the state of their machinery on seven different days. Here is a Naïve Bayes graphical model for the stoplights:

Data:

Day	S_1	S_2	W
1	g	r	w
2	g	r	w
3	g	r	w
4	r	g	w
5	r	g	w
6	r	g	w
7	r	r	b

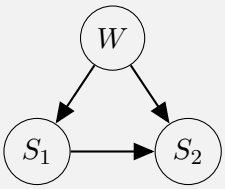
Model:



a. Write the probability tables for $P(W)$, $P(S_1|W)$, $P(S_2|W)$ with the naive Bayes joint distribution that assigns highest probability to the data we observed.

b. What's the posterior probability $P(W = b|S_1 = r, S_2 = r)$?

c. Instead of Naïve Bayes, we use the following graphical model and fill in probability tables with estimates that assign highest probability to the data we observed:



(i) What's the posterior probability $P(W = b|S_1 = r, S_2 = r)$? (Hint: you should not have to do a lot of work.)

(ii) What is it about the problem that makes the second graphical model more suitable than the first?

[]

a.

W	$\mathbf{P}(W)$
w	6/7
b	1/7

S_1	W	$\mathbf{P}(S_1\|W)$
g	w	1/2
r	w	1/2
g	b	0
r	b	1

S_2	W	$\mathbf{P}(S_2\|W)$
g	w	1/2
r	w	1/2
g	b	0
r	b	1

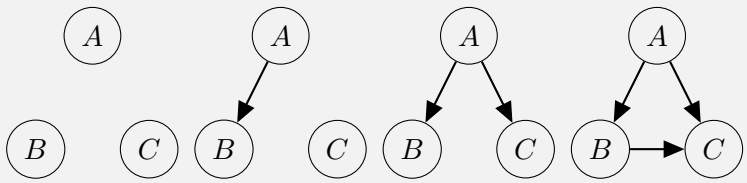
b. 2/5

c. (i) 1

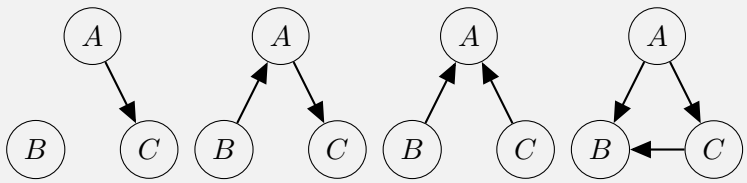
(ii) S_1 and S_2 and not conditionally independent given W in the data (and for real stop lights), since when $W = g$, they are always anti-correlated. The second graphical does not make this assumption, but the first one does.

Exercise 21.XXXX

You want to learn a Bayes' net over the random variables A, B, C. You decide you want to learn not only the Bayes' net parameters, but also the structure from the data. You are willing to consider the 8 structures shown below. First you use your training data to perform maximum likelihood estimation of the parameters of each of the Bayes' nets. Then for each of the learned Bayes' nets, you evaluate the likelihood of the training data (l^{train}), and the likelihood of your cross-validation data (l^{cross}). Both likelihoods are shown below each structure.



	(a)	(b)	(c)	(d)
l^{train}	0.0001	0.0005	0.0015	0.0100
l^{cross}	0.0001	0.0004	0.0011	0.0009



	(e)	(f)	(g)	(h)
l^{train}	0.0008	0.0015	0.0020	0.0100
l^{cross}	0.0006	0.0011	0.0010	0.0009

a. Which Bayes' net structure will (on expectation) perform best on test-data? (If there is a tie, list all Bayes' nets that are tied for the top spot.) Justify your answer.

b. Two pairs of the learned Bayes' nets have identical likelihoods. Explain why this is the case.

c. For every two structures S_1 and S_2, where S_2 can be obtained from S_1 by adding one or more edges, l^{train} is higher for S_2 than for S_1. Explain why this is the case.

[]

a. Bayes' nets (c) and (f), as they have the highest cross validation data likelihood.

b. (c) and (f) have the same likelihoods, and (d) and (h) have the same likelihoods. When learning a Bayes' net with maximum likelihood, we end up selecting the distribution that maximizes the likelihood of the training data from the set of all distributions that can be represented by the Bayes' net structure. (c) and (f) have the same set of conditional independence assumptions, and hence can represent the same set of distributions. This means that they end up with the same distribution as the one that maximizes the training data likelihood, and therefore have identical training and cross validation likelihoods. Same holds true for (d) and (h).

c. When learning a Bayes' net with maximum likelihood, we end up selecting the distribution that maximizes the likelihood of the training data from the set of all distributions that can be represented by the Bayes' net structure. Adding an edge grows the set of distributions that can be represented by the Bayes' net, and can hence only increase the training data likelihood under the best distribution in this set.

Exercise 21.XXXX

The Naïve Bayes model has been famously used for classifying spam. We will use it in the "bag-of-words" model:

- Each email has binary label Y which takes values in $\{\text{spam}, \text{ham}\}$.
- Each word w of an email, no matter where in the email it occurs, is assumed to have probability $P(W = w \mid Y)$, where W takes on words in a pre-determined dictionary. Punctuation is ignored.
- Take an email with K words $w_1, \ldots, w_K$. For instance: email "`hi hi you`" has $w_1 = \text{hi}, w_2 = \text{hi}, w_3 = \text{you}$. Its label is given by:
  $\arg\max_y P(Y = y \mid w_1, \ldots, w_K) = \arg\max_y P(Y = y) \prod_{i=1}^{K} P(W = w_i \mid Y = y)$.

a. You are in possession of a bag of words spam classifier trained on a large corpus of emails. Below is a table of some estimated word probabilities.

W	note	to	self	become	perfect
$P(W \mid Y = \text{spam})$	1/6	1/8	1/4	1/4	1/8
$P(W \mid Y = \text{ham})$	1/8	1/3	1/4	1/12	1/12

You are given a new email to classify, with only two words:

`perfect note`

For what threshold value, c, in the expression $P(Y = \text{spam}) > c$, will the bag of words model predict the label "spam" as the most likely label?

b. You are given only three emails as a training set:

(Spam) `dear sir, I write to you in hope of recovering my gold watch.`

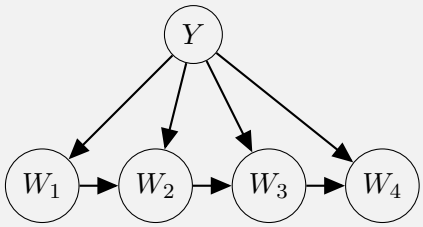
(Ham) `hey, lunch at 12?`
(Ham) `fine, watch it tomorrow night.`

Given the training set, what are the probability estimates for:

(i) $P(W = \text{sir} \mid Y = \text{spam})$
(ii) $P(W = \text{watch} \mid Y = \text{ham})$
(iii) $P(W = \text{gauntlet} \mid Y = \text{ham})$
(iv) $P(Y = \text{ham})$

c. **Becoming less naïve**

We are now going to improve the representational capacity of the model. Presence of word w_i will be modeled not by $P(W = w_i \mid Y)$, where it is only dependent on the label, but by $P(W = w_i \mid Y, W_{i-1})$, where it is also dependent on the previous word. The corresponding model for an email of only four words is given on the right.



(i) With a vocabulary consisting of V words, what is the *minimal* number of conditional **word** probabilities that need to be estimated for this model?

(ii) Which of the following are expected effects of using the new model instead of the old one, if both are trained with a very large set of emails (equal number of spam and ham examples)?

A. The entropy of the posterior $P(Y|W)$ should on average be lower with the new model. (In other words, the model will tend to be more confident in its answers.)
B. The accuracy on the **training** data should be higher with the new model.
C. The accuracy on the **held-out** data should be higher with the new model.

[]

a.

$$P(Y = \text{spam} \mid w_1 = \text{perfect}, w_2 = \text{note}) > P(Y = \text{ham} \mid w_1 = \text{perfect}, w_2 = \text{note})$$
$$\begin{pmatrix} P(w_1 = \text{perfect} \mid Y = \text{s}) \times \\ P(w_2 = \text{note} \mid Y = \text{s}) \times P(Y = \text{s}) \end{pmatrix} > \begin{pmatrix} P(w_1 = \text{perfect} \mid Y = \text{h}) \times \\ P(w_2 = \text{note} \mid Y = \text{h}) \times P(Y = \text{h}) \end{pmatrix}$$
$$1/8 \times 1/6 \times P(Y = \text{spam}) > 1/12 \times 1/8 \times (1 - P(Y = \text{spam}))$$
$$2/96 \times P(Y = \text{spam}) > 1/96 - 1/96 \times P(Y = \text{spam})$$
$$3/96 \times P(Y = \text{spam}) > 1/96$$
$$P(Y = \text{spam}) > 1/3$$

So the threshold is $c = 1/3$

b. (i) Intuitively, the conditional probability is P(word | it's a word in an email of type

Y). We estimate this probability with word counts:

$$P(W = \text{sir} \mid Y = \text{spam}) = \frac{c_w(W = \text{sir}, Y = \text{spam})/c_w(\text{total})}{c_w(Y = \text{spam})/c_w(\text{total})} = \frac{c_w(W = \text{sir}, Y = \text{spam})}{c_w(Y = \text{spam}))} = \frac{1}{13}$$

(ii) Similarly, $P(W = \text{watch} \mid Y = \text{ham}) = \frac{1}{9}$.

(iii) The word "gauntlet" does not occur in our training emails, and since we're not smoothing, we estimate its conditional probability to be 0.

(iv) Estimating the prior probability $P(Y = \text{ham})$ only requires counting emails: $\frac{c_e(Y=\text{ham})}{c_e(\text{total})} = \frac{2}{3}$.

c. (i) We need to consider V possible words for each one of V possible previous words and 2 possible labels: $2V^2$.

(ii) A. False

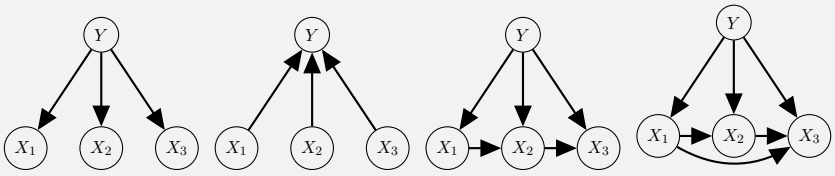
B. True

C. True

The new model is closer to an actual model of language, and so should better model emails and thus filter spam from non-spam on both the training and held-out datasets. Remember that Naïve Bayes is an *overconfident* model: it gives unwarrantedly low-entropy posteriors. This model is less "naïve", and is therefore less overconfident – its posterior can be expected to be higher entropy.

Exercise 21.XXXX

Abhishek has been getting a lot of spam recently(!) and is not satisfied w ith h is email client's Naive Bayes spam classifier. Thankfully he knows about Bayes Nets and has decided to implement his own spam classifier. I t i s y our j ob t o h elp h im m odel h is B ayes N ets and train them. The following are 4 Bayes Nets he is considering. Each variable can take on the values $\{0, 1\}$.



(a) (b) (c) (d)

a. Abhishek wants to know how much memory he needs to store each of these Bayes Nets on disk. The amount of memory depends on the number of values he would need to store in the CPTs for that Bayes Net. For each of the nets above give the **least** number of parameters he would need to store to completely specify the Bayes Net.

b. It's now time to train the Bayes Nets. Abhishek has training datasets D_l, D_m, D_s and a test dataset D_t. The number of training examples in each set vary as $|D_l| > |D_m| > |D_s|$. e_{tr} and e_{te} represent train and test errors and their arguments represent which

dataset a model was trained on. For example, $e_{tr}(A, D_l)$ refers to the training error of model A on dataset D_l.

(i) Abhishek tries a bunch of experiments using model D. In a typical scenario (where train and test data are sampled from the same underlying distribution), order the following errors by their expected value.

A. Order $e_{tr}(D, D_l), e_{tr}(D, D_m), e_{tr}(D, D_s)$
B. Order $e_{te}(D, D_l), e_{te}(D, D_m), e_{te}(D, D_s)$
C. Order $e_{tr}(D, D_l), e_{te}(D, D_l)$
D. Order $e_{tr}(D, D_s), e_{te}(D, D_s)$

(ii) Abhishek is now trying to compare performance across different models. Order the following errors by their expected value:

A. $e_{tr}(A, D_l), e_{tr}(B, D_l), e_{tr}(D, D_l)$
B. $e_{te}(A, D_l), e_{te}(B, D_l), e_{te}(D, D_l)$

[]

a.
- (a): 7
- (b): 11
- (c): 11
- (d): 15

The number of parameters in the Bayes Net is the total number of probability values you need in the CPTs of the Bayes Net. You also need to remember that if A takes k values, $P(A)$ can be represented by $k-1$ values as the last value can be chosen so that the probabilities sum to 1. For example in (A), $P(X_1|Y=0)$ has 1 parameter and so does $P(X_1|Y=1)$ and thus total number of parameters are $2+2+2+1=7$.

b. The rationale is to test knowledge about "power" of a model and overfitting/underfitting. Bayes Nets with less number of parameters have low capacity/power as they can model a more restricted class of distributions due to greater independence assumptions. A more powerful model would overfit on less data and a weak model would underfit. In this question, we are assuming that D_l is *very* large. Rationale for correct answers(column major):

(i) A. $e_{tr}(D, D_l) \geq e_{tr}(D, D_m) \geq e_{tr}(D, D_s)$. A larger dataset would be tougher to model than a smaller. Thus training error would be greater for a larger dataset.

B. $e_{te}(D, D_s) \geq e_{te}(D, D_m) \geq e_{te}(D, D_l)$. A smaller dataset would overfit for a complex model like D and thus testing errors would be more.

C. $e_{tr}(D, D_l) \leq e_{te}(D, D_l)$ Training error less than test error in general because we explicitly model the training data and need to generalize to test data.

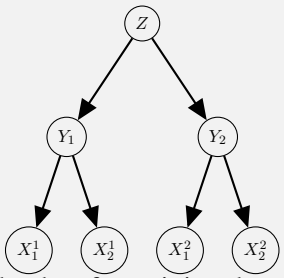
D. $e_{tr}(D, D_s) \leq e_{te}(D, D_s)$ Overfitting for smaller data and complex model

(ii) A. $e_{tr}(A, D_l) \geq e_{tr}(B, D_l) \geq e_{tr}(D, D_l)$. Weaker model underfits on large dataset

B. $e_{te}(A, D_l) \geq e_{te}(B, D_l) \geq e_{te}(D, D_l)$. Same as above. Weaker model cant model the distribution well enough if we have a lot of data.

Exercise 21.XXXX

You are given a model with two distinct label variables Y_1, Y_2, and there is a super label Z which conditions all of these labels, thus giving us this hierarchical naïve Bayes model. The conditional probabilities for the model are parametrized by p_1, p_2, q_0, q_1 and r. **Note that some of the parameters are shared as in the previous part**.



X_1^i	Y_i	$P(X_1^i \mid Y_i)$
0	0	p_1
1	0	$1 - p_1$
0	1	$1 - p_1$
1	1	p_1

X_2^i	Y_i	$P(X_2^i \mid Y_i)$
0	0	p_2
1	0	$1 - p_2$
0	1	$1 - p_2$
1	1	p_2

Y_i	Z	$P(Y_i \mid Z)$
0	0	$1 - q_0$
1	0	q_0
0	1	$1 - q_1$
1	1	q_1

Z	$P(Z)$
0	$1 - r$
1	r

The data for training the model is the following.

sample number	1	2	3	4	5	6	7	8	9	10
X_1^1	0	1	1	0	1	0	1	1	1	0
X_1^2	0	0	0	1	1	1	0	1	0	0
X_2^2	0	0	0	0	0	0	1	0	0	0
X_2^2	0	0	0	0	0	1	0	0	0	0
Y_1	0	0	0	0	1	1	1	1	1	0
Y_2	0	0	0	1	1	1	1	1	0	0
Z	0	0	0	0	1	1	1	1	0	0

a. Compute the maximum likelihood estimate of p_1, p_2, q_0, q_1 and r.

b. Now we are given a partial data point with $X_1^2 = 1, X_2^2 = 1, Y_1 = 1$. Find the probability that $Y_2 = 1$ in terms of the parameters p_1, p_2, q_0, q_1 and r (you might not need all of them).

[]

a. The maximum likelihood estimate of p_1 is the fraction of counts of samples in which $X_1 = Y$. In the given training data, samples $1, 2, 4$ and 6 have $X_1 = Y = 0$ and samples 9 and 10 have $X_1 = Y = 1$, so 6 out of the 10 samples have $X_1 = Y$ and thus $p_1 = \frac{6}{10} = \frac{3}{5}$. Analogously, 8 out of the 10 samples have $X_2 = Y$ and thus $p_2 = \frac{8}{10} = \frac{4}{5}$.

$q_0 = \frac{1}{6}$

$q_1 = 1$

$r = \frac{2}{5}$

b. $P(Y_2 = 1 | X_1^2 = 1, X_2^2 = 1, Y_1 = 1) = \frac{rq_1^2 + (1-r)q_0^2}{rq_1 + (1-r)q_0}$

Exercise 21.BNBX

Explain how to apply the boosting method of Chapter 19 to naive Bayes learning. Test the performance of the resulting algorithm on the restaurant learning problem.

Boosted naive Bayes learning is discussed by Elkan (1997). The application of boosting to naive Bayes is straightforward. The naive Bayes learner uses maximum-likelihood parameter estimation based on counts, so using a weighted training set simply means adding weights rather than counting. Each naive Bayes model is treated as a deterministic classifier that picks the most likely class for each example.

Exercise 21.LINR

Consider N data points (x_j, y_j), where the y_js are generated from the x_js according to the linear Gaussian model in Equation (21.5). Find the values of θ_1, θ_2, and σ that maximize the conditional log likelihood of the data.

We have

$$L = -m(\log\sigma + \log\sqrt{2\pi}) - \sum_j \frac{(y_j - (\theta_1 x_j + \theta_2))^2}{2\sigma^2}$$

hence the equations for the derivatives at the optimum are

$$\begin{aligned}
\frac{\partial L}{\partial \theta_1} &= -\sum_j \frac{x_j(y_j - (\theta_1 x_j + \theta_2))}{\sigma^2} = 0 \\
\frac{\partial L}{\partial \theta_2} &= -\sum_j \frac{(y_j - (\theta_1 x_j + \theta_2))}{\sigma^2} = 0 \\
\frac{\partial L}{\partial \sigma} &= -\frac{m}{\sigma} + \sum_j \frac{(y_j - (\theta_1 x_j + \theta_2))^2}{\sigma^3} = 0
\end{aligned}$$

and the solutions can be computed as

$$\begin{aligned}
\theta_1 &= \frac{m\left(\sum_j x_j y_j\right) - \left(\sum_j y_j\right)\left(\sum_j x_j\right)}{m\left(\sum_j x_j^2\right) - \left(\sum_j x_j\right)^2} \\
\theta_2 &= \frac{1}{m}\sum_j (y_j - \theta_1 x_j) \\
\sigma^2 &= \frac{1}{m}\sum_j (y_j - (\theta_1 x_j + \theta_2))^2
\end{aligned}$$

Exercise 21.XXXX

Answer the following True/False questions.

a. In the case of a binary class and all binary features, Naive Bayes a linear classifier. Justify your answer.

b. Naive Bayes trained using maximum-likelihood parameter estimation is guaranteed not to perform worse if more features are added.

[]

a. $\arg\max_c \Pr(C = c|X_{1:n} = x_{1:n}) = \arg\max_c \log \Pr(C = c, X_{1:n} = x_{1:n})$
We also know $\log \Pr(C = c, X_{1:n} = x_{1:n})$ is linear in $x_{1:n}$.
So this is true.

b. False. If additional dependent features are added, accuracy may be worse.

Exercise 21.MISC.B

Consider the geometric distribution, which has $P(X = k) = (1 - \theta)^{k-1}\theta$. Assume in our training data X took on the values 4, 2, 7, and 9.

a. Write an expression for the log-likelihood of the data as a function of the parameter θ.

b. What is the value of θ that maximizes the log-likelihood, i.e., what is the maximum likelihood estimate for θ?

a. $L(\theta) = P(X = 4)P(X = 2)P(X = 7)P(X = 9)$
$= (1 - \theta)^3\theta(1 - \theta)^1\theta(1 - \theta)^6\theta(1 - \theta)^8\theta$
$= (1 - \theta)^{18}\theta^4$
$\log L(\theta) = 18 \log(1 - \theta) + 4 \log \theta$

b. We take a derivative of the log likelihood equation to find the MLE for theta, θ^{ML}:

- At the maximum we have: $\frac{\partial \log L(\theta)}{\partial \theta} = 18(\frac{-1}{1-\theta}) + 4\frac{1}{\theta} = 0$
- After multiplying both sides by $(1 - \theta)\theta$, $-18\theta + 4(1 - \theta) = 0$ and hence we have an extremum at $\theta = \frac{4}{22}$
- Also: $\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = \frac{-18}{(1-\theta)^2} - \frac{4}{\theta^2} < 0$ hence extremum is indeed a maximum, and hence $\theta^{\mathrm{ML}} = \frac{4}{22}$.

Exercise 21.XXXX

Identical twins are rare, but just how *un*likely are they? With the help of the sociology department, you have a representative sample of twins to help you answer the question. The twins data gives the following observations (a twin refers to one pair of two people):

- m_i = number of identical male twins and f_i = number of identical female twins
- m_f = number of fraternal male twins and f_f = number of fraternal female twins

- b = number of fraternal opposite gender twins

To model this data, we choose these distributions and parameters:

- Twins are identical with probability θ.
- Given identical twins, the twins are male with probability p.
- Given fraternal twins, the probability of male twins is q^2, probability of female twins is $(1-q)^2$ and probability of oppsite gender twins is $2q(1-q)$.

a. Write expressions for the likelihood of the data as a function of the parameters θ, p, and q for the observations $m_i,\ f_i,\ m_f,\ f_f,\ b$.

b. What are the maximum likelihood estimates for θ, p and q?

[]

a. The probability of identical male twins is θp, probability of identical female twins is $\theta(1-p)$, probability of fraternal male twins is $(1-\theta)q^2$, probability of fraternal female twins is $(1-\theta)(1-q)^2$ and probability of fraternal opposite gender twins is $(1-\theta)\cdot 2q(1-q)$. Therefore, the likelihood is

$$\begin{aligned} L(\theta,p,q) &= (\theta p)^{m_i}(\theta(1-p))^{f_i}((1-\theta)q^2)^{m_f}((1-\theta)(1-q)^2)^{f_f}((1-\theta)2q(1-q))^b \\ &= \theta^{(m_i+f_i)}(1-\theta)^{(m_f+f_f+b)}p^{m_i}(1-p)^{f_i}q^{2m_f}(1-q)^{2f_f}(2q(1-q))^b \end{aligned}$$

b. To get the maximum likelihood estimate, we have to maximize the log likelihood by taking partial derivatives,

$$\frac{\partial l}{\partial \theta} = \frac{m_i+f_i}{\theta} - \frac{m_f+f_f+b}{1-\theta} = 0 \qquad \theta_{\text{ML}} = \frac{m_i+f_i}{m_i+f_i+m_f+f_f+b}$$

$$\frac{\partial l}{\partial p} = \frac{m_i}{p} - \frac{m_i+f_i}{1-p} = 0 \qquad p_{\text{ML}} = \frac{m_i}{m_i+f_i}$$

$$\frac{\partial l}{\partial q} = \frac{2m_f+b}{q} - \frac{2f_f+b}{1-q} = 0 \qquad q_{\text{ML}} = \frac{2m_f+b}{2m_f+2f_f+2b}$$

Exercise 21.XXXX

Consider a naive Bayes classifier with two features, shown b elow. We have prior information that the probability model can be parameterized by λ and p, as shown below: Note that $P(X_1=0|Y=0)=P(X_1=1|Y=1)=p$ and $P(X_1|Y)=P(X_2|Y)$ (they share the parameter p). Call this model M1.

We have a training set that contains all of the following:

- n_{000} examples with $X_1 = 0, X_2 = 0, Y = 0$
- n_{010} examples with $X_1 = 0, X_2 = 1, Y = 0$
- n_{100} examples with $X_1 = 1, X_2 = 0, Y = 0$
- n_{110} examples with $X_1 = 1, X_2 = 1, Y = 0$
- n_{001} examples with $X_1 = 0, X_2 = 0, Y = 1$
- n_{011} examples with $X_1 = 0, X_2 = 1, Y = 1$
- n_{101} examples with $X_1 = 1, X_2 = 0, Y = 1$
- n_{111} examples with $X_1 = 1, X_2 = 1, Y = 1$

a. Solve for the maximum likelihood estimate (MLE) of the parameter p with respect to $n_{000}, n_{100}, n_{010}, n_{110}, n_{001}, n_{101}, n_{011}$, and n_{111}.

b. For each of the following values of λ, p, X_1, and X_2, classify the value of Y

(i) $(\lambda = \frac{3}{4}, p = \frac{5}{8}, X_1 = 0, X_2 = 0)$
(ii) $(\lambda = \frac{3}{5}, p = \frac{3}{7}, X_1 = 0, X_2 = 0)$

c. Now let's consider a new model M2, which has the same Bayes' Net structure as M1, but where we have a p_1 value for $P(X_1 = 0 \mid Y = 0) = P(X_1 = 1 \mid Y = 1) = p_1$ and a separate p_2 value for $P(X_2 = 0 \mid Y = 0) = P(X_2 = 1 \mid Y = 1) = p_2$, and we don't constrain $p_1 = p_2$. Let L_{M1} be the likelihood of the training data under model M1 with the maximum likelihood parameters for M1. Let L_{M2} be the likelihood of the training data under model M2 with the maximum likelihood parameters for M2. Are we guaranteed to have $L_{M1} \leq L_{M2}$?

[]

a. We first write down the likelihood of the training data, $T = (Y, X_1, X_1)$.

$$L = \prod_{(y_i, x_{1_i}, x_{2_i})} P(Y = y_i, X_1 = x_{1_i}, X_2 = x_{2_i} | p, \lambda) =$$

$$= \prod_{(y_i, x_{1_i}, x_{2_i})} P(Y = y_i | p, \lambda) P(X_1 = x_{1_i} | p, \lambda) P(X_2 = x_{2_i} | p, \lambda) =$$

$$= \left(\prod_{1}^{n_{000}} \lambda pp \right) \left(\prod_{1}^{n_{001}} (1-\lambda)(1-p)(1-p) \right) \left(\prod_{1}^{n_{010}} \lambda p(1-p) \right) \left(\prod_{1}^{n_{011}} (1-\lambda)(1-p)p \right) *$$

$$* \left(\prod_{1}^{n_{100}} \lambda (1-p)p \right) \left(\prod_{1}^{n_{101}} (1-\lambda)p(1-p) \right) \left(\prod_{1}^{n_{110}} \lambda (1-p)(1-p) \right) \left(\prod_{1}^{n_{111}} (1-\lambda)pp \right) =$$

$$= (\lambda^{n_{000}+n_{010}+n_{100}+n_{110}})((1-\lambda)^{n_{001}+n_{011}+n_{101}+n_{111}})(p^{n_{000}+n_{010}+n_{101}+n_{111}})(p^{n_{000}+n_{011}+n_{100}+n_{111}}) *$$

$$*((1-p)^{n_{001}+n_{011}+n_{100}+n_{110}})((1-p)^{n_{001}+n_{010}+n_{101}+n_{110}}) =$$
$$= (\lambda^{n_{000}+n_{010}+n_{100}+n_{110}})((1-\lambda)^{n_{001}+n_{011}+n_{101}+n_{111}})*$$
$$*(p^{2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}})((1-p)^{2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110}})$$

We then take the logarithm of the likelihood.

$$\log(L) = (n_{000}+n_{010}+n_{100}+n_{110})\log(\lambda)+(n_{001}+n_{011}+n_{101}+n_{111})\log(1-\lambda)+$$

$$+(2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111})\log(p)+(2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110})\log(1-p)$$

We want to take the partial derivative with respect to p and solve for when it is 0 to find the MLE estimate of p. When we do this, the first two terms only depend on λ and not p, so their partial derivative is 0.

$$0 = \frac{\partial}{\partial p}(\log(L)) = (2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111})\frac{1}{p}-(2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110})$$

Multiplying both sides by $(p)(1-p)$, we have:

$$0 = (2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111})(1-p)-(2n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{110})p$$

and simplifying:

$$(2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}) = 2(n_{000}+n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+n_{110}+n_{111})p$$

so

$$p = \frac{2n_{000}+n_{010}+n_{011}+n_{100}+n_{101}+2n_{111}}{2(n_{000}+n_{001}+n_{010}+n_{011}+n_{100}+n_{101}+n_{110}+n_{111})}$$

b. (i) For the first case, $P(Y=0,X_1=0,X_2=0) = \lambda pp$ and $P(Y=1,X_1=0,X_2=0) = (1-\lambda)(1-p)(1-p)$. Since $\lambda > (1-\lambda)$ and $p > (1-p)$, we must have $P(Y=0,X_1=0,X_2=0) > (Y=1,X_1=0,X_2=0)$.

For the second case, we have $P(Y=0,X_1=1,X_2=0) = \lambda(1-p)p$ and $P(Y=1,X_1=1,X_2=0) = (1-\lambda)p(1-p)$. Since both expressions have a $p(1-p)$ term, the question is reduced to $\lambda < (1-\lambda)$ so $P(Y=1,X_1=1,X_2=0) > P(Y=0,X_1=1,X_2=0)$.

Thus the predicted label is $Y=1$

(ii) $P(Y=0,X_1=0,X_2=0) = \lambda pp = \frac{3}{5}\frac{3}{7}\frac{3}{7} = \frac{27}{5*7*7}$

We also know $P(Y=1,X_1=0,X_2=0) = (1-\lambda)(1-p)(1-p) = \frac{2}{5}\frac{4}{7}\frac{4}{7} = \frac{32}{5*7*7}$.

So $\frac{27}{5*7*7} = P(Y=0,X_1=0,X_2=0) < P(Y=1,X_1=0,X_2=0) = \frac{32}{5*7*7}$.

Thus the predicted label is $Y=1$.

c. $M2$ can represent all of the same probability distributions that $M1$ can, but also some

more (when $p_1 \neq p_2$). So in general $M2$ allows for more fitting of the training data, which results in a higher likelihood.

Exercise 21.NORX

Consider the noisy-OR model for fever described in Section 13.2.2. Explain how to apply maximum-likelihood learning to fit the parameters of such a model to a set of complete data. (*Hint*: use the chain rule for partial derivatives.)

There are a couple of ways to solve this problem. Here, we show the indicator vari-able method described on page 794. Assume we have a child variable Y with parents $X_1, \ldots, X_k$ and let the range of each variable be $\{0, 1\}$. Let the noisy-OR parameters be $q_i = P(Y = 0|X_i = 1, X_{-i} = 0)$. The noisy-OR model then asserts that

$$P(Y = 1|x_1, \ldots, x_k) = 1 - \prod_{i=1}^{k} q_i^{x_i} .$$

Assume we have m complete-data samples with values y_j for Y and x_{ij} for each X_i. The conditional log likelihood for $P(Y|X_1, \ldots, X_k)$ is given by

$$\begin{aligned} L &= \sum_j \log \left(1 - \prod_i q_i^{x_i j}\right)^{y_j} \left(\prod_i q_i^{x_i j}\right)^{1-y_j} \\ &= \sum_j y_j \log \left(1 - \prod_i q_i^{x_i j}\right) + (1 - y_j) \sum_i x_{ij} \log q_i \end{aligned}$$

The gradient with respect to each noisy-OR parameter is

$$\begin{aligned} \frac{\partial L}{\partial q_i} &= \sum_j -\frac{y_j x_{ij} \prod_i q_i^{x_i j}}{q_i \left(1 - \prod_i q_i^{x_i j}\right)} + \frac{(1 - y_j) x_{ij}}{q_i} \\ &= \sum_j \frac{x_{ij} \left(1 - y_j - \prod_i q_i^{x_i j}\right)}{q_i \left(1 - \prod_i q_i^{x_i j}\right)} \end{aligned}$$

Exercise 21.BETI

This exercise investigates properties of the Beta distribution defined in Equation (21.6).

a. By integrating over the range $[0, 1]$, show that the normalization constant for the distribution $Beta(\theta; a, b)$ is given by $\alpha = \Gamma(a + b)/\Gamma(a)\Gamma(b)$ where $\Gamma(x)$ is the **Gamma function**, defined by $\Gamma(x + 1) = x \cdot \Gamma(x)$ and $\Gamma(1) = 1$. (For integer x, $\Gamma(x + 1) = x!$.)

b. Show that the mean is $a/(a + b)$.

c. Find the mode(s) (the most likely value(s) of θ).

d. Describe the distribution $Beta(; \epsilon, \epsilon)$ for very small ϵ. What happens as such a distribution is updated?

a. We will solve this for positive integer a and b by induction over a. Let $\alpha(a, b)$ be the normalization constant. For the base cases, we have

$$\alpha(1, b) = 1/\int_0^1 \theta^0(1-\theta)^{b-1}d\theta = -1/[\frac{1}{b}(1-\theta)^b]_0^1 = b$$

and

$$\frac{\Gamma(1+b)}{\Gamma(1)\Gamma(b)} = \frac{b \cdot \Gamma(b)}{1 \cdot \Gamma(b)} = b\,.$$

For the inductive step, we assume for all b that

$$\alpha(a-1, b+1) = \frac{\Gamma(a+b)}{\Gamma(a-1)\Gamma(b+1)} = \frac{a-1}{b} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

Now we evaluate $\alpha(a, b)$ using integration by parts. We have

$$\begin{aligned} 1/\alpha(a,b) &= \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta \\ &= [\theta^{a-1} \cdot \frac{1}{b}(1-\theta)^b]_1^0 + \frac{a-1}{b}\int_0^1 \theta^{a-2}(1-\theta)^b d\theta \\ &= 0 + \frac{a-1}{b}\frac{1}{\alpha(a-1, b+1)} \end{aligned}$$

Hence

$$\alpha(a, b) = \frac{b}{a-1}\alpha(a-1, b+1) = \frac{b}{a-1}\frac{a-1}{b} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

as required.

b. The mean is given by the following integral:

$$\begin{aligned} \mu(a,b) &= \alpha(a,b)\int_0^1 \theta \cdot \theta^{a-1}(1-\theta)^{b-1}d\theta \\ &= \alpha(a,b)\int_0^1 \theta^a(1-\theta)^{b-1}d\theta \\ &= \alpha(a,b)/\alpha(a+1,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b+1)} = \frac{a}{a+b}\,. \end{aligned}$$

c. The mode is found by solving for $dBeta(\theta; a, b)/d\theta = 0$:

$$\begin{aligned}
&\frac{d}{d\theta}(\alpha(a,b)\theta^{a-1}(1-\theta)^{b-1}) \\
&= \alpha(a,b)[(a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2}] = 0 \\
&\Rightarrow (a-1)(1-\theta) = (b-1)\theta \\
&\Rightarrow \theta = \frac{a-1}{a+b-2}
\end{aligned}$$

d. $Beta(\theta; \epsilon, \epsilon) = \alpha(\epsilon, \epsilon)\theta^{\epsilon-1}(1-\theta)^{\epsilon-1}$ tends to very large values close to $\theta = 0$ and $\theta = 1$, i.e., it expresses the prior belief that the distribution characterized by θ is nearly deterministic (either positively or negatively). After updating with a positive example we obtain the distribution $Beta(; 1+\epsilon, \epsilon)$, which has nearly all its mass near $\theta = 1$ (and the converse for a negative example), i.e., we have learned that the distribution characterized by θ is deterministic in the positive sense. If we see a "counterexample", e.g., a positive and a negative example, we obtain $Beta(\theta; 1+\epsilon, 1+\epsilon)$, which is close to uniform, i.e., the hypothesis of near-determinism is abandoned.

Exercise 21.MLPA

Consider an arbitrary Bayesian network, a complete data set for that network, and the likelihood for the data set according to the network. Give a simple proof that the likelihood of the data cannot decrease if we add a new link to the network and recompute the maximum-likelihood parameter values.

Consider the maximum-likelihood parameter values for the CPT of node Y in the original network, where an extra parent X_{k+1} will be added to Y. If we set the parameters for $P(y|x_1, \ldots, x_k, x_{k+1})$ in the new network to be identical to $P(y|x_1, \ldots, x_k)$ in the original network, regardless of the value x_{k+1}, then the likelihood of the data is unchanged. Maximizing the likelihood by altering the parameters can then only *increase* the likelihood.

Exercise 21.LIKR

Consider a single Boolean random variable Y (the "classification"). Let the prior probability $P(Y = true)$ be π. Let's try to find π, given a training set $D = (y_1, \ldots, y_N)$ with N independent samples of Y. Furthermore, suppose p of the N are positive and n of the N are negative.

a. Write down an expression for the likelihood of D (i.e., the probability of seeing this particular sequence of examples, given a fixed value of π) in terms of π, p, and n.

b. By differentiating the log likelihood L, find the value of π that maximizes the likelihood.

c. Now suppose we add in k Boolean random variables $X_1, X_2, \ldots, X_k$ (the "attributes") that describe each sample, and suppose we assume that the attributes are conditionally independent of each other given the goal Y. Draw the Bayes net corresponding to this assumption.

d. Write down the likelihood for the data including the attributes, using the following additional notation:

- α_i is $P(X_i = true|Y = true)$.
- β_i is $P(X_i = true|Y = false)$.
- p_i^+ is the count of samples for which $X_i = true$ and $Y = true$.
- n_i^+ is the count of samples for which $X_i = false$ and $Y = true$.
- p_i^- is the count of samples for which $X_i = true$ and $Y = false$.
- n_i^- is the count of samples for which $X_i = false$ and $Y = false$.

[*Hint*: consider first the probability of seeing a single example with specified values for $X_1, X_2, \ldots, X_k$ and Y.]

e. By differentiating the log likelihood L, find the values of α_i and β_i (in terms of the various counts) that maximize the likelihood and say in words what these values represent.

f. Let $k = 2$, and consider a data set with 4 all four possible examples of the XOR function. Compute the maximum likelihood estimates of π, α_1, α_2, β_1, and β_2.

g. Given these estimates of π, α_1, α_2, β_1, and β_2, what are the posterior probabilities $P(Y = true|x_1, x_2)$ for each example?

a. The probability of a positive example is π and of a negative example is $(1 - \pi)$, and the data are independent, so the probability of the data is $\pi^p(1 - \pi)^n$

b. We have $L = p \log \pi + n \log(1 - \pi)$; if the derivative is zero, we have

$$\frac{\partial L}{\partial \pi} = \frac{p}{\pi} - \frac{n}{1 - \pi} = 0$$

so the ML value is $\pi = p/(p + n)$, i.e., the proportion of positive examples in the data.

c. This is the "naive Bayes" probability model.

d. The likelihood of a single instance is a product of terms. For a positive example, π times α_i for each true attribute and $(1 - \alpha_i)$ for each negative attribute; for a negative example, $(1-\pi)$ times β_i for each true attribute and $(1-\beta_i)$ for each negative attribute. Over the whole data set, the likelihood is $\pi^p(1-\pi)^n \prod_i \alpha_i^{p_i^+} (1-\alpha_i)^{n_i^+} \beta_i^{p_i^-} (1-\beta_i)^{n_i^-}$.

e. The log likelihood is

$L = p \log \pi + n \log(1-\pi) + \sum_i p_i^+ \log \alpha_i + n_i^+ \log(1-\alpha_i) + p_i^- \log \beta_i + n_i^- \log(1-\beta_i)$.

Setting the derivatives w.r.t. α_i and β_i to zero, we have

$$\frac{\partial L}{\partial \alpha_i} = \frac{p_i^+}{\alpha_i} - \frac{n_i^+}{1 - \alpha_i} = 0 \quad \text{and} \quad \frac{\partial L}{\partial \beta_i} = \frac{p_i^-}{\beta_i} - \frac{n_i^-}{1 - \beta_i} = 0$$

giving $\alpha_i = p_i^+/(p_i^+ + n_i^+)$, i.e., the fraction of cases where X_i is true given Y is true, and $\beta_i = p_i^-/(p_i^- + n_i^-)$, i.e., the fraction of cases where X_i is true given Y is false.

f. In the data set we have $p = 2$, $n = 2$, $p_i^+ = 1$, $n_i^+ = 1$, $p_i^- = 1$, $n_i^- = 1$. From our formulæ, we obtain $\pi = \alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 0.5$.

g. Each example is predicted to be positive with probability 0.5.

21.3 Learning with Hidden Variables: The EM Algorithm

Exercise 21.EMLE

Consider the application of EM to learn the parameters for the network in Figure 21.14(a), given the true parameters in Equation (21.9).

a. Explain why the EM algorithm would not work if there were just two attributes in the model rather than three.

b. Show the calculations for the first iteration of EM starting from Equation (21.10).

c. What happens if we start with all the parameters set to the same value p? (*Hint*: you may find it helpful to investigate this empirically before deriving the general result.)

d. Write out an expression for the log likelihood of the tabulated candy data on page 793 in terms of the parameters, calculate the partial derivatives with respect to each parameter, and investigate the nature of the fixed point reached in part (c).

a. Consider the ideal case in which the bags were infinitely large so there is no statistical fluctuation in the sample. With two attributes (say, *Flavor* and *Wrapper*), we have five unknowns: θ gives the the relative sizes of the bags, θ_{F1} and θ_{F2} give the proportion of cherry candies in each bag, and θ_{W1} and θ_{W2} give the proportion of red wrappers in each bag. In the data, we observe just the flavor and wrapper for each candy; there are four combinations, so three independent numbers can be obtained. This is not enough to recover five unknowns. With three attributes, there are eight combinations and seven numbers can be obtained, enough to recover the seven parameters.

b. The computation for $\theta^{(1)}$ has eight nearly identical expressions and calculations, one of which is shown. The symbolic expression for $\theta_{F1}^{(1)}$ is shown, but not its evaluation; it would be reasonable to ask students to write out the expression in terms of the parameters, as was done for $\theta^{(1)}$, and calculate the value. The final answers are given in the chapter.

c. Consider the contribution to the update for θ from the 273 red-wrapped cherry candies with holes:

$$\frac{273}{1000} \cdot \frac{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)}}{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)} + \theta_{F2}^{(0)}\theta_{W2}^{(0)}\theta_{H2}^{(0)}(1-\theta^{(0)})}$$

If all of the seven named parameters have value p, this reduces to

$$\frac{273}{1000} \cdot \frac{p^4}{p^4 + p^3(1-p)} = \frac{273p}{1000}$$

with similar results for the other candy categories. Thus, the new value for $\theta^{(1)}$ just ends up being $1000p/1000 = p$.

We can check the expression for θ_{F1} too; for example, the 273 red-wrapped cherry candies with holes contribute an expected count of

$$\begin{aligned} & 273P(Bag = 1 \mid Flavor_j = cherry, Wrapper = red, Holes = 1) \\ &= 273\frac{\theta_{F1}\theta_{W1}\theta_{H1}\theta}{\theta_{F1}\theta_{W1}\theta_{H1}\theta + \theta_{F2}\theta_{W2}\theta_{H2}(1-\theta)} = 273p \end{aligned}$$

and the 90 green-wrapped cherry candies with no holes contribute an expected count of

$$\begin{aligned} & 90P(Bag = 1 \mid Flavor_j = cherry, Wrapper = green, Holes = 0) \\ &= 90\frac{\theta_{F1}(1-\theta_{W1})(1-\theta_{H1})\theta}{\theta_{F1}(1-\theta_{W1})(1-\theta_{H1})\theta + \theta_{F2}(1-\theta_{W2})(1-\theta_{H2})(1-\theta)} \\ &= 90p^2(1-p)^2/p(1-p)^2 = 90p\ . \end{aligned}$$

Continuing, we find that the new value for θ_{F1} is $560p/1000p = 0.56$, the proportion of cherry candies in the entire sample.

For θ_{F2}, the 273 red-wrapped cherry candies with holes contribute an expected count of

$$\begin{aligned} & 273P(Bag = 2 \mid Flavor_j = cherry, Wrapper = red, Holes = 1) \\ &= 273\frac{\theta_{F2}\theta_{W2}\theta_{H2}(1-\theta)}{\theta_{F1}\theta_{W1}\theta_{H1}\theta + \theta_{F2}\theta_{W2}\theta_{H2}(1-\theta)} = 273(1-p) \end{aligned}$$

with similar contributions from the other cherry categories, so the new value is $560(1-p)/1000(1-p) = 0.56$, as for θ_{F1}. Similarly, $\theta^{(1)}_{W1} = \theta^{(1)}_{W2} = 0.545$, the proportion of red wrappers in the sample, and $\theta^{(1)}_{H1} = \theta^{(1)}_{H2} = 0.550$, the proportion of candies with holes in the sample.

Intuitively, this makes sense: because the bag label is invisible, labels 1 and 2 are *a priori* indistinguishable; initializing all the conditional parameters to the same value (regardless of the bag) provides no means of breaking the symmetry. Thus, the symmetry remains.

On the next iteration, we no longer have all the parameters set to p, but we do know that, for example,

$$\theta_{F1}\theta_{W1}\theta_{H1} = \theta_{F2}\theta_{W2}\theta_{H2}$$

so those terms cancel top and bottom in the expression for the contribution of the 273 candies to θ_{F1}, and once again the contribution is $273p$.

To cut a long story short, all the parameters remain fixed after the first iteration, with θ at its initial value p and the other parameters at the corresponding empirical frequencies as indicated above.

d. This part takes some time but makes the abstract mathematical expressions in the chap-

ter very concrete! The one concession to abstraction will be the use of symbols for the empirical counts, e.g.,

$$N_{cr1} = N(Flavor = cherry, Wrapper = red, Holes = 1) = 273 \ .$$

with marginal counts N_c, N_{r1}, etc. Thus we have $\theta_{F1}^{(1)} = N_c/N = 560/1000$.

The log likelihood is given by

$$\begin{aligned} L(\mathbf{d}) &= \log P(\mathbf{d}) = \log \prod_j P(d_j) = \sum_j \log P(d_j) \\ &= N_{cr1} \log P(F = cherry, W = red, H = 1) + \\ &\quad N_{lr1} \log P(F = lime, W = red, H = 1) + \\ &\quad N_{cr0} \log P(F = cherry, W = red, H = 0) + \\ &\quad N_{lr0} \log P(F = lime, W = red, H = 0) + \\ &\quad N_{cg1} \log P(F = cherry, W = green, H = 1) + \\ &\quad N_{lg1} \log P(F = lime, W = green, H = 1) + \\ &\quad N_{cg0} \log P(F = cherry, W = green, H = 0) + \\ &\quad N_{lg0} \log P(F = lime, W = green, H = 0) \end{aligned}$$

Each of these probabilities can be expressed in terms of the network parameters, giving the following expression for $L(\mathbf{d})$:

$$\begin{aligned} &N_{cr1} \log(\theta_{F1}\theta_{W1}\theta_{H1}\theta + \theta_{F2}\theta_{W2}\theta_{H2}(1-\theta)) + \\ &N_{lr1} \log((1-\theta_{F1})\theta_{W1}\theta_{H1}\theta + (1-\theta_{F2})\theta_{W2}\theta_{H2}(1-\theta)) + \\ &N_{cr0} \log(\theta_{F1}\theta_{W1}(1-\theta_{H1})\theta + \theta_{F2}\theta_{W2}(1-\theta_{H2})(1-\theta)) + \\ &N_{lr0} \log((1-\theta_{F1})\theta_{W1}(1-\theta_{H1})\theta + (1-\theta_{F2})\theta_{W2}(1-\theta_{H2})(1-\theta)) + \\ &N_{cg1} \log(\theta_{F1}(1-\theta_{W1})\theta_{H1}\theta + \theta_{F2}(1-\theta_{W2})\theta_{H2}(1-\theta)) + \\ &N_{lg1} \log((1-\theta_{F1})(1-\theta_{W1})\theta_{H1}\theta + (1-\theta_{F2})(1-\theta_{W2})\theta_{H2}(1-\theta)) + \\ &N_{cg0} \log(\theta_{F1}(1-\theta_{W1})(1-\theta_{H1})\theta + \theta_{F2}(1-\theta_{W2})(1-\theta_{H2})(1-\theta)) + \\ &N_{lg0} \log((1-\theta_{F1})(1-\theta_{W1})(1-\theta_{H1})\theta + (1-\theta_{F2})(1-\theta_{W2})(1-\theta_{H2})(1-\theta)) \end{aligned}$$

Hence $\partial L/\partial\theta$ is given by

$$
\begin{aligned}
&N_{cr1}\,\frac{\theta_{F1}\theta_{W1}\theta_{H1} - \theta_{F2}\theta_{W2}\theta_{H2}}{\theta_{F1}\theta_{W1}\theta_{H1}\theta + \theta_{F2}\theta_{W2}\theta_{H2}(1-\theta)} \\
&-N_{lr1}\,\frac{(1-\theta_{F1})\theta_{W1}\theta_{H1} - (1-\theta_{F2})\theta_{W2}\theta_{H2}}{(1-\theta_{F1})\theta_{W1}\theta_{H1}\theta + (1-\theta_{F2})\theta_{W2}\theta_{H2}(1-\theta)} \\
&+N_{cr0}\,\frac{\theta_{F1}\theta_{W1}(1-\theta_{H1}) - \theta_{F2}\theta_{W2}(1-\theta_{H2})}{\theta_{F1}\theta_{W1}(1-\theta_{H1})\theta + \theta_{F2}\theta_{W2}(1-\theta_{H2})(1-\theta)} \\
&-N_{lr0}\,\frac{(1-\theta_{F1})\theta_{W1}(1-\theta_{H1}) - (1-\theta_{F2})\theta_{W2}(1-\theta_{H2})}{(1-\theta_{F1})\theta_{W1}(1-\theta_{H1})\theta + (1-\theta_{F2})\theta_{W2}(1-\theta_{H2})(1-\theta)} \\
&+N_{cg1}\,\frac{\theta_{F1}(1-\theta_{W1})\theta_{H1} - \theta_{F2}(1-\theta_{W2})\theta_{H2}}{\theta_{F1}(1-\theta_{W1})\theta_{H1}\theta + \theta_{F2}(1-\theta_{W2})\theta_{H2}(1-\theta)} \\
&-N_{lg1}\,\frac{(1-\theta_{F1})(1-\theta_{W1})\theta_{H1} - (1-\theta_{F2})(1-\theta_{W2})\theta_{H2}}{(1-\theta_{F1})(1-\theta_{W1})\theta_{H1}\theta + (1-\theta_{F2})(1-\theta_{W2})\theta_{H2}(1-\theta)} \\
&+N_{cg0}\,\frac{\theta_{F1}(1-\theta_{W1})(1-\theta_{H1}) - \theta_{F2}(1-\theta_{W2})(1-\theta_{H2})}{\theta_{F1}(1-\theta_{W1})(1-\theta_{H1})\theta + \theta_{F2}(1-\theta_{W2})(1-\theta_{H2})(1-\theta)} \\
&-N_{lg0}\,\frac{(1-\theta_{F1})(1-\theta_{W1})(1-\theta_{H1}) - (1-\theta_{F2})(1-\theta_{W2})(1-\theta_{H2})}{(1-\theta_{F1})(1-\theta_{W1})(1-\theta_{H1})\theta + (1-\theta_{F2})(1-\theta_{W2})(1-\theta_{H2})(1-\theta)}
\end{aligned}
$$

By inspection, we can see that whenever $\theta_{F1} = \theta_{F2}$, $\theta_{W1} = \theta_{W2}$, and $\theta_{H1} = \theta_{H2}$, the derivative is identically zero. Moreover, each term in the above expression has the form $k/f(\theta)$ where k does not contain θ and $f'(\theta)$ evaluates to zero under these conditions. Thus the second derivative $\partial^2 L/\partial\theta^2$ is a collection of terms of the form $-kf'(\theta)/(f(\theta))^2$, all of which evaluate to zero. In fact, all derivatives evaluate to zero under these conditions, so the likelihood is completely flat with respect to θ in the subspace defined by $\theta_{F1} = \theta_{F2}$, $\theta_{W1} = \theta_{W2}$, and $\theta_{H1} = \theta_{H2}$. Another way to see this is to note that, in this subspace, the terms within the logs in the expression for $L(\mathbf{d})$ simplify to terms of the form $\phi_F\phi_W\phi_H\theta + \phi_F\phi_W\phi_H(1-\theta) = \phi_F\phi_W\phi_H$, so that the likelihood is in fact independent of θ!

A representative partial derivative $\partial L/\partial\theta_{F1}$ is given by

$$
\begin{aligned}
&N_{cr1}\,\frac{\theta_{W1}\theta_{H1}\theta}{\theta_{F1}\theta_{W1}\theta_{H1}\theta+\theta_{F2}\theta_{W2}\theta_{H2}(1-\theta)}\\
&-N_{lr1}\,\frac{\theta_{W1}\theta_{H1}\theta}{(1-\theta_{F1})\theta_{W1}\theta_{H1}\theta+(1-\theta_{F2})\theta_{W2}\theta_{H2}(1-\theta)}\\
&+N_{cr0}\,\frac{\theta_{W1}(1-\theta_{H1})\theta}{\theta_{F1}\theta_{W1}(1-\theta_{H1})\theta+\theta_{F2}\theta_{W2}(1-\theta_{H2})(1-\theta)}\\
&-N_{lr0}\,\frac{\theta_{W1}(1-\theta_{H1})\theta}{(1-\theta_{F1})\theta_{W1}(1-\theta_{H1})\theta+(1-\theta_{F2})\theta_{W2}(1-\theta_{H2})(1-\theta)}\\
&+N_{cg1}\,\frac{(1-\theta_{W1})\theta_{H1}\theta}{\theta_{F1}(1-\theta_{W1})\theta_{H1}\theta+\theta_{F2}(1-\theta_{W2})\theta_{H2}(1-\theta)}\\
&-N_{lg1}\,\frac{(1-\theta_{W1})\theta_{H1}\theta}{(1-\theta_{F1})(1-\theta_{W1})\theta_{H1}\theta+(1-\theta_{F2})(1-\theta_{W2})\theta_{H2}(1-\theta)}\\
&+N_{cg0}\,\frac{(1-\theta_{W1})(1-\theta_{H1})\theta}{\theta_{F1}(1-\theta_{W1})(1-\theta_{H1})\theta+\theta_{F2}(1-\theta_{W2})(1-\theta_{H2})(1-\theta)}\\
&-N_{lg0}\,\frac{(1-\theta_{W1})(1-\theta_{H1})\theta}{(1-\theta_{F1})(1-\theta_{W1})(1-\theta_{H1})\theta+(1-\theta_{F2})(1-\theta_{W2})(1-\theta_{H2})(1-\theta)}
\end{aligned}
$$

Unlike the previous case, here the individual terms do not evaluate to zero. Writing $\theta_{F1}=\theta_{F2}=N_c/N$, etc., the expression for $\partial L/\partial\theta_{F1}$ becomes

$$
\begin{aligned}
&N_{cr1}\,\frac{N\,N_rN_1\theta}{N_cN_rN_1\theta+N_cN_rN_1(1-\theta)}\\
&-N_{lr1}\,\frac{N\,N_rN_1\theta}{(N-N_c)N_rN_1\theta+(N-N_c)N_rN_1(1-\theta)}\\
&+N_{cr0}\,\frac{N\,N_r(N-N_1)\theta}{N_cN_r(N-N_1)\theta+N_cN_r(N-N_1)(1-\theta)}\\
&-N_{lr0}\,\frac{N\,N_r(N-N_1)\theta}{(N-N_c)N_r(N-N_1)\theta+(N-N_c)N_r(N-N_1)(1-\theta)}\\
&+N_{cg1}\,\frac{N\,(N-N_r)N_1\theta}{N_c(N-N_r)N_1\theta+N_c(N-N_r)N_1(1-\theta)}\\
&-N_{lg1}\,\frac{N\,(N-N_r)N_1\theta}{(N-N_c)(N-N_r)N_1\theta+(N-N_c)(N-N_r)N_1(1-\theta)}\\
&+N_{cg0}\,\frac{N\,(N-N_r)(N-N_1)\theta}{N_c(N-N_r)(N-N_1)\theta+N_c(N-N_r)(N-N_1)(1-\theta)}\\
&-N_{lg0}\,\frac{N\,(N-N_r)(N-N_1)\theta}{(N-N_c)(N-N_r)(N-N_1)\theta+(N-N_c)(N-N_r)(N-N_1)(1-\theta)}
\end{aligned}
$$

This in turn simplifies to

$$
\begin{aligned}
\frac{\partial L}{\partial\theta_{F1}} &= \frac{(N_{cr1}+N_{cr0}+N_{cg1}+N_{cg0})N\theta}{N_c}-\frac{(N_{lr1}+N_{lr0}+N_{lg1}+N_{lg0})N\theta}{N-N_c}\\
&= \frac{N_cN\theta}{N_c}-\frac{(N-N_c)N\theta}{N-N_c}=0\,.
\end{aligned}
$$

Thus, we have a stationary point as expected.

To identify the nature of the stationary point, we need to examine the second derivatives. We will not do this exhaustively, but will note that

$$\begin{aligned}\partial^2 L/\partial\theta_{F1}^2 \;=\; & -N_{cr1}\,\frac{(\theta_{W1}\theta_{H1}\theta)^2}{(\theta_{F1}\theta_{W1}\theta_{H1}\theta+\theta_{F2}\theta_{W2}\theta_{H2}(1-\theta))^2} \\ & -N_{lr1}\,\frac{(\theta_{W1}\theta_{H1}\theta)^2}{((1-\theta_{F1})\theta_{W1}\theta_{H1}\theta+(1-\theta_{F2})\theta_{W2}\theta_{H2}(1-\theta))^2}\cdots\end{aligned}$$

with all terms negative, suggesting (possibly) a local maximum in the likelihood surface. A full analysis requires evaluating the Hessian matrix of second derivatives and calculating its eigenvalues.

[[need exercises]]