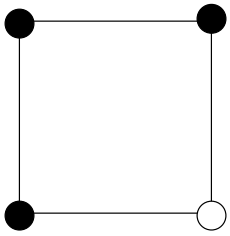
CS 530 — Principles of AI
Written Exercises
Out: October 2, 2003
Due: October 14, 2003

The following figure represents a decision rule for classifying an observation of two binary features X and Y into two classes, c_1 and c_2.



Each vertex represents a possible observation: the horizontal axis indicates the value of feature X; the vertical axis indicates the value of feature Y. A black circle at that vertex represents a decision to classify the observation as class c_1 and a white one represents a decision to classify it as class c_2.

Problem 1. State the four constraints on the probability distributions of features and classes which are required for this decision rule to yield the lowest possible error.

Problem 2. Is this decision rule compatible with a representation as a linear classifier? In other words, are the region where the classifier reports class c_1 and the region where the classifer reports class c_2 separable by a hyperplane?

Problem 3. Construct a specific joint probability distribution on variables X, Y and C that leads to this decision rule and which satisfies the Naive Bayes assumption: that X and Y are conditionally independent given C. *Hint:* You can simply "estimate" parameters for your Naive Bayes model assuming the four training "samples" illustrated by the decision rule itself.

Problem 4. Now consider the following joint probability distribution on X, Y and C:

$$\begin{array}{ll} P(X=1,Y=1,C=c_1)=3/20 & P(X=1,Y=1,C=c_2)=2/20 \\ P(X=1,Y=0,C=c_1)=0/20 & P(X=1,Y=0,C=c_2)=5/20 \\ P(X=0,Y=1,C=c_1)=3/20 & P(X=0,Y=1,C=c_2)=2/20 \\ P(X=0,Y=0,C=c_1)=3/20 & P(X=0,Y=0,C=c_2)=2/20 \end{array}$$

In this case the features are *not* conditionally independent given the class. Nevertheless, the joint probability distribution still leads to the decision rule illustrated above. (Check this for yourself.)

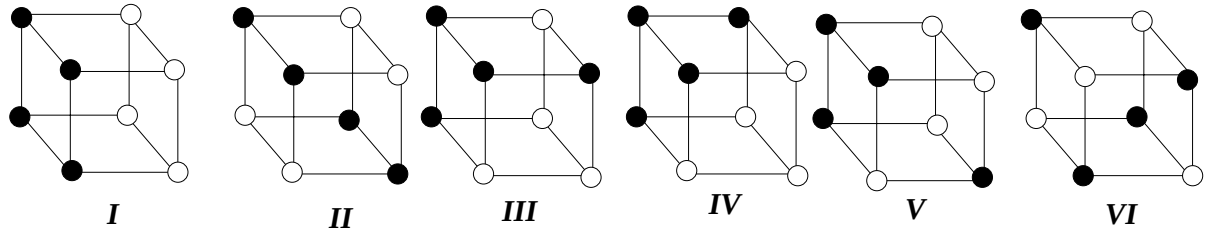
Suppose you have a large amount of training data from this distribution. You can therefore assume that the eight kinds of event above occur in the training data with exactly the specified probabilities. You can use this data to estimate parameters for a Naive Bayes model by a statistical method. Of course, the Naive Bayes model will not describe the distribution exactly, since its assumptions are not met—but all assumptions are only approximate anyway.

What parameter values do you estimate in this case? What decision rule does the Naive Bayes model of the distribution give? (In calculating the decision rule, you will have the opportunity to explore the relationships between $P(X,Y,C)$ and $P(C|X,Y)=P(X,Y,C)/P(X,Y)$ in a classification setting—which may be helpful for you if you're still rusty with these issues.)

Problem 5. In general, it is possible to distinguish between two kinds of learning procedures. One kind of learning assumes that the world meets certain assumptions, and uses statistical methods to estimate parameters correctly when those assumptions hold. Another kind of learning assumes that the rule for optimal performance instantiates a certain format, and searches for the best rule in this class by optimization.

In a couple of brief sentences, discuss the what examples like that of Problem 3 and 4 have to say about these two kinds of learners, for the case of Naive Bayes modeling versus optimizing a linear decision boundary. (Actually, a number of important pattern-recognition algorithms are based on optimizing a linear decision boundary, including the Perceptron and the Support-Vector Machine.)

Problem 6. It turns out that—ignoring the labels and axes for the features and the labels of the classes—there are only six qualitatively different rules possible for splitting the eight possible observations of three binary features into two categories each covering four members. The possibilities are shown below.

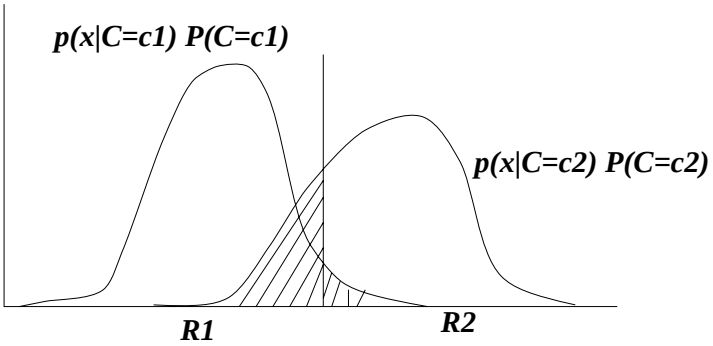


In which cases can the decision boundary be understood as a hyperplane? In which cases can the decision boundary not be represented as a hyperplane? Accordingly, in which cases is it possible to find a distribution underlying the classification rule in which the feature values are conditionally independent given the class?

Problem 7. Again, ignoring the labels and axes for the features and the labels for the classes, there are *two* qualitatively different rules for splitting the four possible observations of two binary features into two categories each covering two members. Draw the possibilities.

What evidence does the pattern suggested by Problems 6 and 7 give about the following question: if you know that features are *not* conditionally independent given the class, how reasonable is it to expect that the best decision boundary will take a linear form?

Problem 8. In class, I used a diagram of the following sort to illustrate the probability of error for Bayesian classifier deciding between $C = c_1$ and $C = c_2$ on the basis of a noisy measurement x.



Show a corresponding overlaid graph of $P(C = c_1|x)$ and $P(C = c_2|x)$. You can't read the probability of error of a classifier off of your new graph. In at most two sentences, explain why not.

Problem 9. Consider the following decision rule for a two-category one-dimensional problem: decide $C = c_1$ if $x > \theta$; otherwise decide $C = c_2$.

(a) Show that the probability of error for this rule is given by

$$P(\text{error}) = P(C = c_1) \int_{-\infty}^{\theta} p(x|C = c_1)dx + P(C = c_2) \int_{\theta}^{\infty} p(x|C = c_2)dx$$

(b) By differentating, show that a necessary condition to minimize $P(\text{error})$ is that θ satisfy

$$p(\theta|C = c_1)P(C = c_1) = p(\theta|C = c_2)P(C = c_2)$$

(c) Does this equation define θ uniquely?

(d) Give an example where a value of θ satisfying the equation actually *maximizes* the probability of error.

Problem 10. Continuous density estimation can suffer from the same fragility as discrete density estimation. If the assumptions of the model are wrong, the model can lead you to do something very stupid. Alternatively, the best strategy for using the model might require parameter values that are very different from the way the world actually is.

This example brings out the theme. Suppose we have two equally probable categories (i.e., $P(C = c_1) = P(C = c_2) = 0.5$). Further, we know that $p(x|C = c_1) \sim N(0,1)$ but *assume* that $p(x|C = c_2) \sim N(\mu,1)$. (Thus, μ is the parameter we seek to estimate.) Imagine however that the *true* underlying distribution is $p(x|C = c_2) \sim N(1,10^6)$.

(a) Given training data, we can use "the method of maximum likelihood" (an easy statistical technique) to estimate μ: we use the mean of the data as $\hat{\mu}$. What value do we expect for $\hat{\mu}$ given an indefinitely large amount of data?

(b) What is the decision boundary arising from this maximum likelihood estimate in the poor model?

(c) You can use the equation from Problem 9b to derive a correct decision boundary given the *true* underlying distributions— $p(x|C = c_1) \sim N(0.1)$ and $p(x|C = c_2) \sim N(1,10^6)$. How should we classify an observation x? Be careful to include all portions of the decision boundary.

(d) Now consider again classifiers based on the (poor) model assumption $p(x|C = c_2) \sim N(\mu,1)$. Using your result immediately above, find a *new* value of μ that will give lower error than a classifier that is trained statistically.