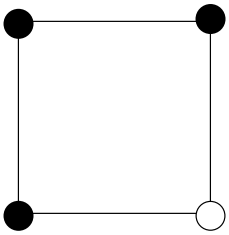
CS 530 — Principles of AI
Written Exercises
Out: October 2, 2001
Due: October 18, 2001

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

Problem 1b. 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 1c. 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 1d. 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.)

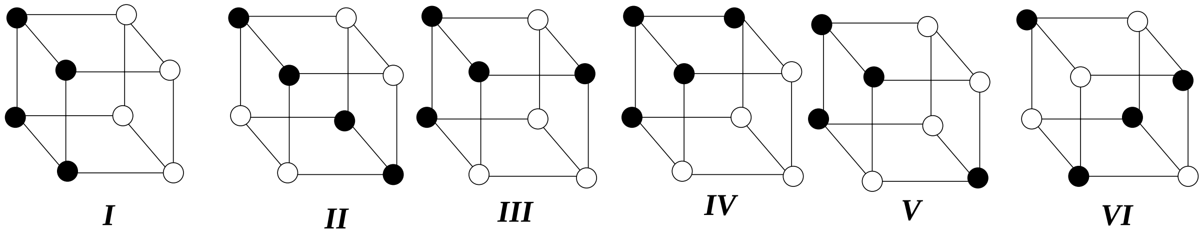
Suppose you have a large amount of training data from this distribution. You can therefore assume that the eight events above occur in the training data with exactly the specified probabilities. You can use this data to estimate parameters for a Naive Bayes model by a statistical method. Of course, the 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 1e. 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 1c and 1d 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; see Chapter 1.5 and Chapters 3.4–3.5 in Bishop's text.)

Problem 1f. 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.

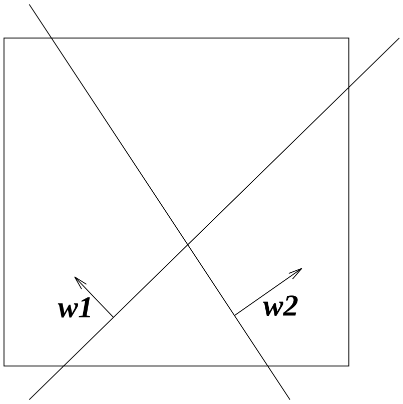


In which cases can the decision boundary be understood as a hyperplane? In which cases can the decision boundary not be represented as a hyperplane? Accordingly, in which cases is it possible to find a distribution underlying the classification rule in which the feature values are conditionally independent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

Problem 1g. 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.

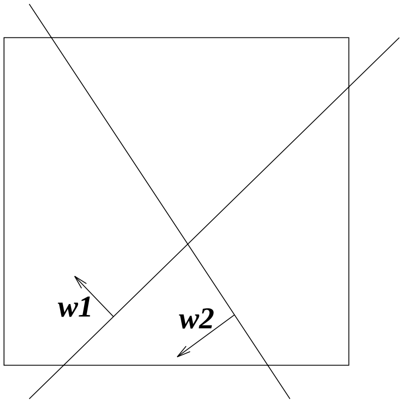
What evidence does the pattern suggested by Problems 1g and 1f give about the following question: if you know that features are *not* conditionally independent given the class, how reasonable is it to expect that the best decision boundary will take a linear form?

Problem 2a. We've seen that a Naive Bayes model assigns log probability of class membership proportional to the distance above a plane corresponding to the class; the normal to the plane is a weight vector, as in the diagram below (where the box represents a feature space, weight vector $w1$ corresponds to class c_1 and weight vector $w2$ corresponds to class c_2).



Use a diagram (and an explanatory sentence) to indicate geometrically how these planes determine a plane decision boundary for a Naive Bayes classifier (with decision regions on either side).

Problem 2b. Now do the same for this case.



That is, use a diagram to indicate geometrically how these planes determine a plane decision boundary for a Naive Bayes classifier (with decision regions on either side). What is different?