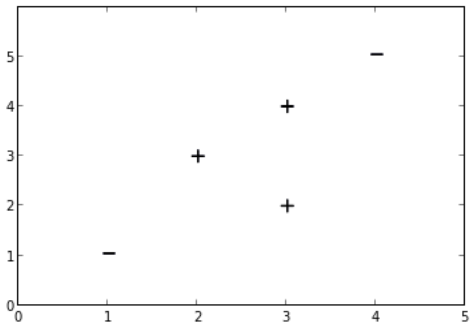
CS188 Spring 2013 Section 9: Machine Learning

You want to predict if movies will be profitable based on their screenplays. You hire two critics A and B to read a script you have and rate it on a scale of 1 to 5. The critics are not perfect; here are five data points including the critics' scores and the performance of the movie:

Movie Name	A	B	Profit?
Pellet Power	1	1	No
Ghosts!	3	2	Yes
Pac is Bac	4	5	No
Not a Pizza	3	4	Yes
Endless Maze	2	3	Yes

Training Data

First, you would like to examine the linear separability of the data. Plot the data on the 2D plane below; label profitable movies with + and non-profitable movies with − and determine if the data are linearly separable.



The data are not linearly separable!

Now you first decide to use a perceptron to classify your data. This problem will use the multi-class formulation even though there are only two classes. Suppose you directly use the scores given above as features, together with a bias feature. That is $f_0 = 1$, $f_1 =$ score given by A and $f_2 =$ score given by B.

1. You want to train the perceptron on the training data in Table 1. The initial weights are given below:

Profit	Weights	Weights after 1st update
Yes	[-1, 0, 0]	[0, 3, 2]
No	[1, 0, 0]	[0, -3, -2]

   (i) Which is the first training instance at which you update your weights? "Ghosts!", because the perceptron predicts "No" for profit, whereas the label is "Yes."

   (ii) In the table above, write the updated weights after the first update. Note that in the binary case, the weights for two classes are exactly opposites, which is why we usually simplify and just work with a single weight vector.

2. More generally, irrespective of the training data, you want to know if your features are powerful enough to allow you to handle a range of scenarios. Some scenarios are given on the next page. Circle those scenarios for which a perceptron using the features above can indeed perfectly classify the data.

   (i) Your reviewers are awesome: if the total of their scores is more than 8, then the movie will definitely be a success and otherwise it will fail. Can classify (consider weights $[-8, 1, 1]$)

   (ii) Your reviewers are art critics. Your movie will succeed if and only if each reviewer gives either a score of 2 or a score of 3. Cannot classify

   (iii) Your reviewers have weird but different tastes. Your movie will succeed if and only if both reviewers agree. Cannot classify

You decide to use a different set of features. Consider the following feature space:

$$
\begin{aligned}
f_0 &= 1 \text{ (The bias feature)} \\
f_{1A} &= 1 \text{ if score given by A is 1, 0 otherwise} \\
f_{1B} &= 1 \text{ if score given by B is 1, 0 otherwise} \\
f_{2A} &= 1 \text{ if score given by A is 2, 0 otherwise} \\
f_{2B} &= 1 \text{ if score given by B is 2, 0 otherwise} \\
&\dots \\
f_{5B} &= 1 \text{ if score given by B is 5, 0 otherwise}
\end{aligned}
$$

3. Consider again the three scenarios in part 2. Using a perceptron with the new features, which of the three scenarios can be perfectly classified? Circle your answer(s) below:

   (i) Your reviewers are awesome: if the total of their scores is more than 8, then the movie will definitely be a success, and otherwise it will fail. Can classify (consider weights $[-8, 1, 1, 2, 2, \dots, 5, 5]$)

   (ii) Your reviewers are art critics. Your movie will succeed if and only if each reviewer gives either a score of 2 or a score of 3. Can classify (consider weights $[1, -\infty, -\infty, 0, 0, 0, 0, -\infty, -\infty, -\infty, \infty]$, or $[-1.5, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0]$)

   (iii) Your reviewers have weird but different tastes. Your movie will succeed if and only if both reviewers agree. Cannot classify

You have just heard of naive Bayes and you want to use a naive Bayes classifier. You use the scores given by the reviewers as the features of the naive Bayes classifier, i.e., the random variables in your naive Bayes model are A and B, each with a domain of $\{1, 2, \dots, 5\}$, and *Profit* with a domain of *Yes* and *No*.

4. Draw the Bayes net corresponding to the naive Bayes model on the back of this page.

5. List the types of the conditional probability tables you need to estimate along with their sizes (e.g., $P(X \mid Y)$ has 24 entries).

Probability	Size
$P(Profit)$	2
$P(A\|Profit)$	10
$P(B\|Profit)$	10

6. Your nephew is taking the CS188 class at Berkeley. He claims that the naive Bayes classifier you just built is actually a linear classifier in the feature space used for part 3. In other words, the decision boundary of the naive Bayes classifier is a hyperplane in this feature space. For the positive class, what is the weight of the feature f_{3B} in terms of the parameters of the naive Bayes model? You can answer in symbols, but be precise. (Hint: Consider the log of the probability.)

   The weight is $\log P(B = 3|Profit = Yes)$.

   Consider the following weight vectors:

$$w_{Yes} = [\log P(Profit = Yes), \log P(A = 1|Profit = Yes), \ldots, \log P(B = 5|Profit = Yes)]$$

$$w_{No} = [\log P(Profit = No), \log P(A = 1|Profit = No), \ldots, \log P(B = 5|Profit = No)].$$

   Using these weight vectors, the linear classification rule

$$y = \arg\max_y w_y \cdot f(x)$$

   is equivalent to the Naive Bayes rule

$$y = \arg\max_y P(Profit = y)P(A = f_A(x)|Profit = y)P(B = f_B(x)|Profit = y).$$

   This is because the weights are log probabilities, so summing the weights is equivalent to multiplying probabilities, and the 0/1 features pick out which entries of the conditional probability table to include in the product.