AI534 Written Homework Assignment 3

Woonki Kim

kimwoon@oregonstate.edu

1. (Naive Bayes Classifier) (7 pts) Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

(a) (3 pts) Learn a Naive Bayes classifier by estimating all necessary probabilities (there should be 7 independent probabilities to be estimated in total).

1. Learn prior $P(Y = j)$ for $j = 1, 2$:

$$P(Y = 0) = \frac{1}{2}, P(Y = 1) = \frac{1}{2}$$

2.Learn $P(A, B, C|y = 0)$

- Class condition for $y = 0$

$$(A = 0) : \frac{2}{3}, (A = 1) : \frac{1}{3}$$

$$(B = 0) : \frac{1}{3}, (B = 1) : \frac{2}{3}$$

$$(C = 0) : \frac{1}{3}, (C = 1) : \frac{2}{3}$$

3.Learn $P(A, B, C|y = 1)$

- Class condition for $y = 1$

$$(A = 0) : \frac{1}{3}, (A = 1) : \frac{2}{3}$$

$$(B = 0) : \frac{1}{3}, (B = 1) : \frac{2}{3}$$

$$(C = 0) : \frac{2}{3}, (C = 1) : \frac{1}{3}$$

(b) (3 pts) Compute the probability $P(y = 1|A = 1, B = 0, C = 0)$.

$$P(y = 1|A = 1, B = 0, C = 0)$$

$$= \frac{\prod P(A = 1, B = 0, C = 0|Y = 1)P(Y)}{\sum_{j=1}^{2} \prod P(A = 1, B = 0, C = 0|Y = j) \times P(Y = j)}$$

$$= \frac{P(A = 1|Y = 1)P(B = 0|Y = 1)P(C = 0|Y = 1)P(Y = 1)}{\sum_{j=1}^{2} P(A = 1|Y = j)P(B = 0|Y = j)P(C = 0|Y = j) \times P(y = j)}$$

$$= \frac{\frac{2}{3}\frac{1}{3}\frac{2}{3} \times \frac{1}{2}}{\frac{2}{3}\frac{1}{3}\frac{2}{3} \times \frac{1}{2} + \frac{1}{3}\frac{1}{3}\frac{1}{3} \times \frac{1}{2}}$$

$$= \frac{4}{5}$$

(c) (1 pts) Suppose we know that the three features A, B and C are independent from one another, can we say that the Naive Bayes assumption is valid? (Note that the particular data set is irrelevant for this question). If your answer is yes, please explain why; if you answer is no please give an counter example.

If A,B and C are independent from one another, Naive Bayes assumption is invalid.

Naive Bayes assumption : x is conditionally independent of y given z, if $\forall i, j, k \quad P(x = i, y = j, z = k) = P(x = i \mid z = k)P(y = j \mid z = k)$
or equivalently $p(x|y, z) = p(x|z)$ *or* $p(x, y|z) = p(x|z)p(y|z)$

For example, to make Naive Bayes assumption valid $P(A = 0, B = 0|Y = 0) = P(A = 0 \mid Y = 0)P(B = 0 \mid Y = 0)$ *should hold.*

While, if three features are independent to each other, $P(A = 0, B = 0|Y = 0) = P(A = 0) \times P(B = 0)$

Let's say for independent features A and B
$P(A = 0) = 0.5, P(B = 0) = 0.5, P(Y = 0) = 0.5$
$P(A = 0|Y = 0) = 0.8, P(B = 0|Y = 0) = 0.9$

Naive Bayes assumption, $P(A = 0, B = 0|Y = 0) = P(A = 0 \mid Y = 0)P(B = 0 \mid Y = 0)$

While, $P(A = 0, B = 0|Y = 0) = 0.9, P(A = 0) \times P(B = 0) = 0.64$
Showing that Naive Bayes is invalid for independent features.

2. (Naive Bayes learns linear decision boundary.) (10 pts) Show that the following naive Bayes classifiers learn linear decision boundary $w_0 + w_1x_1 + w_2x_2 + ... + w_dx_d = 0$. Express the weights using the corresponding Naive Bayes parameters. Hint: start with the decision boundary defined by $\log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = 0$.

(a) Bernoulli Naive Bayes model, where features $x_1, x_2, ..., x_d$ are binary indicating the presence/absence of words in the vocabulary.

$$\log \frac{P(y = 1|\mathbf{x})}{P(y = 0|\mathbf{x})} = 0$$

$$P(y = 1 \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y = 1)P(y = 1)}{P(\mathbf{x})}, \quad P(y = 0 \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y = 0)P(y = 0)}{P(\mathbf{x})}$$

$$\log \frac{P(y = 1 \mid \mathbf{x})}{P(y = 0 \mid \mathbf{x})} = \log \frac{P(\mathbf{x} \mid y = 1)P(y = 1)}{P(\mathbf{x} \mid y = 0)P(y = 0)} = \log P(\mathbf{x} \mid y = 1)P(y = 1) - \log P(\mathbf{x} \mid y = 0)P(y = 0) = 0$$

Thus,

$$\log P(\mathbf{x} \mid y = 1) + \log P(y = 1) - \log P(\mathbf{x} \mid y = 0) - \log P(y = 0) = 0$$

While by Naive Bayes assumption,

$$P(\mathbf{x} \mid y) = \prod_{i=1}^{n} P(x_i \mid y)$$

Substituting into formula:

$$\log \prod_{i=1}^{n} P(x_i \mid y = 1) + \log P(y = 0) - \log \prod_{i=1}^{n} P(x_i \mid y = 0) - \log P(y = 1) = 0$$

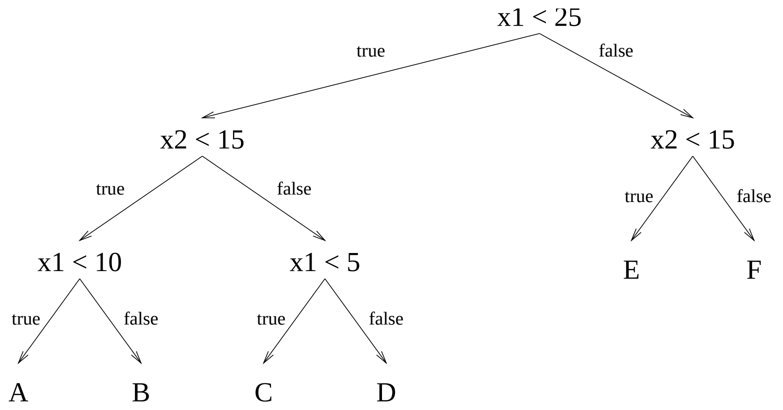
$$\Rightarrow \sum_{i=1}^{n} (\log P(x_i \mid y = 1) - \log P(x_i \mid y = 0)) + \log P(y = 0) - \log P(y = 1) = 0$$

While for Bernoulli $P(x_i|y)$ and $\log P(x_i|y)$ is linear in x_i.
We can express the formula as:

$$w_0 + \sum_{i=1}^{n} w_i x_i = 0, \ \textit{where } w_0 = \log P(y = 0) - \log P(y = 1), w_i = \log P(x_i \mid y = 1) - \log P(x_i \mid y = 0)$$

(b) Multinomial Naive Bayes Model, where $x_1, ..., x_d$ representing counts of words $w_1, ..., w_d$ in the vocabulary. Express the weights using the Naive Bayes parameters: the class priors $P(y = 1), P(y = 0)$ and the class conditionals: $p(w_i|y = 1)andp(w_i|y = 0)$

3. (6 pts) Consider the following decision tree:



(a) (2 pts) Draw the decision boundaries defined by this tree. Each leaf of the tree is labeled with a letter. Write this letter in the corresponding region of input space.

See Figure 1.

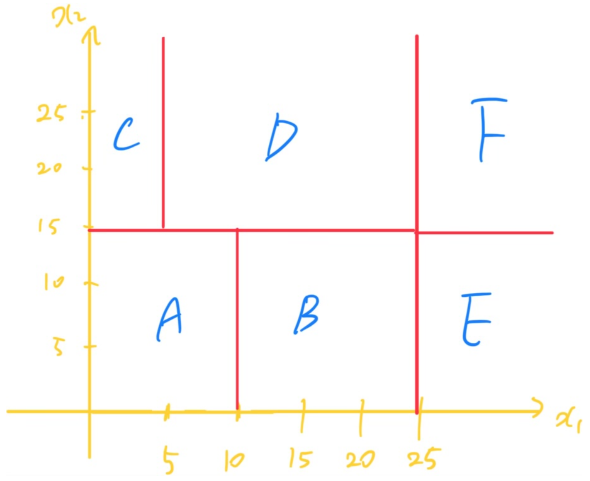


Figure 1: A descriptive caption for Figure 1.

(b) (2 pts) Give another decision tree that is syntactically different but defines the same decision boundaries. This demonstrates that the space of decision trees is syntactically redundant.

See Figure 2.

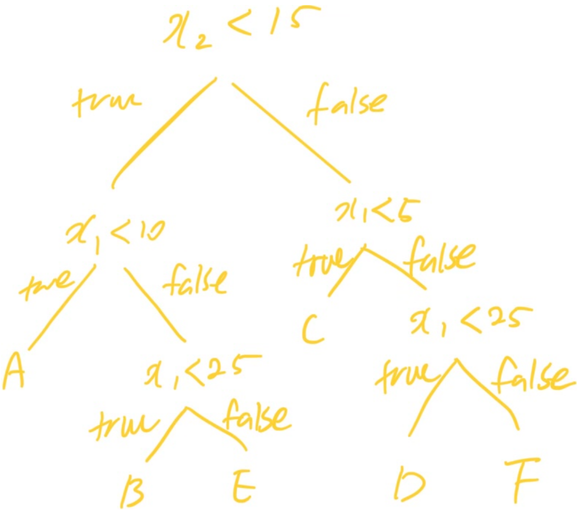


Figure 2:

(c) (2pts) How does this redundancy influence learning (does it make it easier or harder to find an accurate tree)?

It will make harder to find an accurate tree.

By this redundancy in decision tree space, it makes us to consider more possibilities to find an optimal decision tree compared to when the tree is not redundant.

4. (6 pts) In the basic decision tree algorithm (assuming we always create binary splits), we choose the feature/value pair with the maximum information gain as the test to use at each internal node of the decision tree. Suppose we modified the algorithm to choose at random from among those feature/value combinations that had non-zero mutual information, and we kept all other parts of the algorithm unchanged.

(a) (2 pts) What is the maximum number of leaf nodes that such a decision tree could contain if it were trained on m training examples?

Maximum achievable leaf node is m when having m training examples.

(b) (2 pts) What is the maximum number of leaf nodes that a decision tree could contain if it were trained on m training examples using the original maximum mutual information version of the algorithm? Is it bigger, smaller, or the same as your answer to (b)?

It is going to be same, since original algoritm would grow tree until every training examples are divided.

(c) (2 pts)How do you think this change (using random splits vs. maximum information mutual information splits) would affect the testing accuracy of the decision trees produced on average? Why?

Randomly choosing features would result in low testing accuracy. The reason why we choose the maximum information gain is to focus on more discriminative features, making the decision tree to be more general so that it fits better on to unseen data. But if we choose random features, it will increase both underfitting and overfitting possibilities resulting in poor testing possibilities.

5. (8 pts) Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

Learn a decision tree from the training set shown above using the information gain criterion. Show your steps, including the calculation of information gain (you can skip $H(y)$ and just compute $H(y|\mathbf{x})$) of different candidate tests. You can randomly break ties (or better, choose the one that give you smaller tree if you do a bit look ahead for this problem).

Reusing estimated probabilities solved in problem 1.

-Prior

$$P(Y=0)=\frac{1}{2}, P(Y=1)=\frac{1}{2}$$

- Class condition for $y=0$

$$(A=0):\frac{2}{3},(A=1):\frac{1}{3}$$

$$(B=0):\frac{1}{3},(B=1):\frac{2}{3}$$

$$(C=0):\frac{1}{3},(C=1):\frac{2}{3}$$

3.Learn $P(A,B,C|y=1)$

- Class condition for $y=1$

$$(A=0):\frac{1}{3},(A=1):\frac{2}{3}$$

$$(B=0):\frac{1}{3},(B=1):\frac{2}{3}$$

$$(C=0):\frac{2}{3},(C=1):\frac{1}{3}$$

Split on A:
Entropy for $A=0$:

$$H(Y\mid A=0)=-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}$$

Entropy for $A=1$:

$$H(Y\mid A=1)=-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}$$

Weighted Entropy:

$$H(Y\mid A)=-\frac{1}{2}(\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})-\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})=-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}$$

Split on B:
Entropy for $B=0$:

$$H(Y\mid B=0)=-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}$$

Entropy for $B = 1$:

$$H(Y \mid B = 1) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

Weighted Entropy:

$$H(Y \mid B) = -\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) - \frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

Split on C:
Entropy for $C = 0$:

$$H(Y \mid C = 0) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

Entropy for $C = 1$:

$$H(Y \mid C = 0) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$$

Weighted Entropy:

$$H(Y \mid C) = -\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) - \frac{1}{2}(\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

Combined uncertainty is same for every split, thus mutual information would be same as well.
It won't matter what to split first in this case.
But if we see closer look at the table, we could easily figure out that dividing B first would reduce tree size. And when $B = 1$, if $C = 1Y = 0$ follows. Similarly if $C = 0, Y = 1$. So root is B, and when $B = 1$ next node is C. Below is my tree according to this explanation.

See Figure 3.

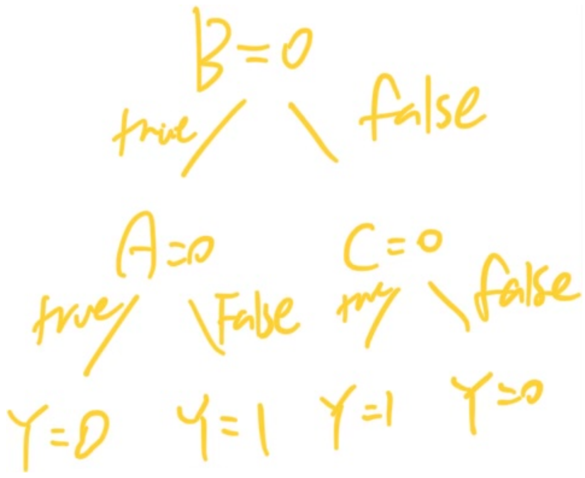


Figure 3:

6. (8pts) Prove that

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

.

Hint: you should start with the definition $H(X, Y) = -\sum_{x,y} P(x, y) \log P(x, y)$. Here we use X, Y to denote the random variables and x, y denote the values X and Y take, $P(x, y)$ is a short hand notation denoting $P(X = x, Y = y)$.

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y),$$

1. The joint entropy $H(X, Y)$:

$$H(X, Y) = -\sum_{x,y} P(x, y) \log P(x, y).$$

2. Expand Using the chain rule of probabilities:

$$H(X, Y) = -\sum_{x,y} P(x, y) \log P(x, y) = -\sum_{x,y} P(x, y) \log[P(x)P(y \mid x)].$$

Put log over it:

$$H(X, Y) = -\sum_{x,y} P(x, y)[\log P(x) + \log P(y \mid x)].$$

$$H(X, Y) = -\sum_{x,y} P(x, y) \log P(x) - \sum_{x,y} P(x, y) \log P(y \mid x).$$

3. Now looking at the first term, we can marginallize over y which is equal to $H(X)$:

$$-\sum_{x,y} P(x, y) \log P(x) = -\sum_{x} P(x) \log P(x) = H(X).$$

4. In the second term, by definition of conditional probability which is equal to $H(Y|X)$:

$$-\sum_{x,y} P(x, y) \log P(y \mid x) = -\sum_{x} P(x) \sum_{y} P(y \mid x) \log P(y \mid x) = H(Y \mid X)$$

Thus,

$$H(X, Y) = H(X) + H(Y \mid X).$$

6. Do same process for $P(x, y) = P(y)P(x \mid y)$:

$$H(X, Y) = H(Y) + H(X \mid Y).$$

7. Final result:

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y).$$