10-315 Introduction to Machine Learning (SCS Majors)
Lectures 5: Naive Bayes

Leila Wehbe
Carnegie Mellon University
Machine Learning Department

Reading: <a href ="http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf"> (http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf">) Generative and Disciminative Classifiers </a> by Tom Mitchell.

Lecture outcomes:

- Conditional Independence
- Naïve Bayes, Gaussian Naive Bayes
- Practical Examples

Assume you want to build a classifier for new customers

O	I	S	J	Y
Is older than 35 years	Has Personal Income	Is a Student	Birthday before July 1st	Buys computer
0	1	0	0	0
0	1	0	1	0
1	1	0	1	1
1	1	0	1	1
1	0	1	0	1
1	0	1	1	0
0	0	1	0	1
0	1	0	0	0
0	0	1	1	1
0	1	1	0	1
0	0	1	1	1
1	1	0	1	1
0	1	1	0	1
1	1	0	0	0

What to predict for the next customer? We want to find
$P(Y = 0|O = 0, I = 0, S = 1, J = 1)$

O	I	S	J	Y
Is older than 35 years	Has Personal Income	Is a Student	Birthday before July 1st	Buys computer
0	1	1	1	?

How many parameters must we estimate?

Suppose $X = (X_1, X_2)$ where X_i and Y are boolean random variables

How many parameters do we need to estimate to know $P(Y|X1, X2)$?

How many parameters must we estimate?

Suppose $X = (X_1, X_2)$ where X_i and Y are boolean random variables

How many parameters do we need to estimate to know $P(Y|X1, X2)$?

X_1	X_2	$P(Y = 1 \mid X_1, X_2)$	$P(Y = 0 \mid X_1, X_2)$
0	1	0.1	0.9
1	0	0.24	0.76
0	1	0.54	0.46
1	1	0.23	0.77

How many parameters must we estimate?

Suppose $X = (X_1, X_2)$ where X_i and Y are boolean random variables

How many parameters do we need to estimate to know $P(Y|X1, X2)$?

X_1	X_2	$P(Y = 1 \mid X_1, X_2)$	$P(Y = 0 \mid X_1, X_2)$
0	1		0.9
1	0		0.76
0	1		0.46
1	1		0.77

4 parameters: $P(Y = 0 \mid X_1, X_2) = 1 - P(Y = 1 \mid X_1, X_2)$

How many parameters must we estimate?

Suppose $X = (X_1, X_2, \ldots, X_d)$ where X_i and Y are boolean random variables

How many parameters do we need to estimate to know $P(Y|X1, \ldots X_d)$?

X_1	X_2	X_3	…	X_d	$P(Y = 1 \mid X)$	$P(Y = 0 \mid X)$
0	0	0	…	0	0.1	0.9
0	0	0	…	1	0.24	0.76
…	…	…	…	…	…	…
…	…	…	…	…	…	…
1	1	1	…	1	0.52	0.48

How many parameters must we estimate?

Suppose $X = (X_1, X_2, \ldots, X_d)$ where X_i and Y are boolean random variables

How many parameters do we need to estimate to know $P(Y|X_1, \ldots X_d)$?

X_1	X_2	X_3	...	X_d	$P(Y = 1 \mid X)$	$P(Y = 0 \mid X)$
0	0	0	...	0	0.1	0.9
0	0	0	...	1	0.24	0.76
...	...	...	...	...	...	...
...	...	...	...	...	...	...
1	1	1	...	1	0.52	0.48

2^d rows! (2^d parameters!).

If we have 30 boolean X_i's: ~ 1 billion rows!

Last Lecture we asked:

Can we just estimate P(Y|X) in this fashion and be done?

We might not have enough data. For example consider having 100 attributes of people:

- how many rows will we have? $2^{100} > 10^{30}$
- how many people on earth? 10^{10}
- 99.99\% of rows will not have training examples!

Can we use Bayes Rule to reduce the number of parameters?

We used bayes rule last lecture to express marginal and conditional probabilities of the data and parameter θ:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

We can also use it for the conditional distribution of Y given X:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

BTW, this notation is a shorthand for:

$$(\forall i, j)\ P(Y = y_i|X = x_i) = \frac{P(X = x_i|Y = y_i)P(Y = y_i)}{P(X = x_i)}$$

Can we use Bayes Rule to reduce the number of parameters?

We used bayes rule last lecture to express marginal and conditional probabilities of the data and parameter θ:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

We can also use it for the conditional distribution of Y given X:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

BTW, this notation is a shorthand for:

$$(\forall i, j)\ P(Y = y_i|X = x_i) = \frac{P(X = x_i|Y = y_i)P(Y = y_i)}{P(X = x_i)}$$

Equivalently:

$$(\forall i, j)\ P(Y = y_i|X = x_i) = \frac{P(X = x_i|Y = y_i)P(Y = y_i)}{\sum_k P(X = x_i|Y = y_k)P(Y = y_k)}$$

Does Bayes Rule help reduce the number of parameters?

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X)$ might be expensive to estimate, but it's the same for both classes, so we might be able to avoid computing it.
- $P(Y)$ is only 1 parameter.
- What about $P(X|Y)$? i.e. $P(X_1, X_2, \ldots X_d)$. How many parameters do we need?

How many parameters do we need for $P(X_1, X_2, \ldots X_d)$?

Y	X_1	X_2	X_3	…	X_d	$P(X \mid Y)$
0	0	0	0	…	0	
0	0	0	0	…	1	
0	…	…	…	…	…	
0	1	1	1	…	1	
1	0	0	0	…	0	
1	0	0	0	…	1	
1	…	…	…	…	…	
1	1	1	1	…	1	

- How many parameters for the red cells?
- Should I compute parameters for the blue cells as well?

How many parameters do we need for $P(X_1, X_2, \ldots X_d)$?

Y	X_1	X_2	X_3	…	X_d	$P(X \mid Y)$
0	0	0	0	…	0	
0	0	0	0	…	1	
0	…	…	…	…	…	
0	1	1	1	…	1	
1	0	0	0	…	0	
1	0	0	0	…	1	
1	…	…	…	…	…	
1	1	1	1	…	1	

- How many parameters for the red cells? $2^d - 1$
- Should I compute parameters for the blue cells as well? yes, $2^d - 1$

Does Bayes Rule help reduce the number of parameters?

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X)$ is the same for both classes, so we might be able to avoid computing it (also, I can get it for free if I know $P(X|Y)$ and $P(Y)$).
- $P(Y)$ is only 1 parameter.
- $P(X_1, X_2, \ldots X_d|Y)$: $2(2^d - 1)$ parameters

Still too many parameters!

If we have 30 boolean X_i's: ~ 2 billion!

Solution:

1- Be smart about how to estimate probabilities from sparse data

- maximum likelihood estimates
- maximum a posteriori estimates
- Be smart about how to represent joint distributions
- Bayes networks, graphical models, conditional independence

Be smart about how to represent joint distributions

Conditional Independence:

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z.

$$(\forall i, j, k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_j)$$

Which we often write:

$$P(X|Y, Z) = P(X|Z)$$

For example: $P(\text{thunder}|\text{rain, lightning}) = P(\text{thunder}|\text{lightning})$

- Thunder is independent of rain **given lightning**.
    - Once we know there if there is or there is lightning, no more information is provided by the value of rain.
- **This does not mean that thunder is independent of rain**.

The Naïve Bayes Algorithm

Naïve Bayes is a classifier that assumes conditional independence of the variables X_i given the label Y. For example: $P(X_1|X_2, Y) = P(X_1|Y)$

How does this assumption simplify $P(X_1, X_2|Y)$?

The Naïve Bayes Algorithm

Naïve Bayes is a classifier that assumes conditional independence of the variables X_i given the label Y. For example: $P(X_1|X_2, Y) = P(X_1|Y)$

How does this assumption simplify $P(X_1, X_2|Y)$?

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2, Y) \qquad \text{(chain rule)}$$

The Naïve Bayes Algorithm

Naïve Bayes is a classifier that assumes conditional independence of the variables X_i given the label Y. For example: $P(X_1|X_2, Y) = P(X_1|Y)$

How does this assumption simplify $P(X_1, X_2|Y)$?

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2, Y) && \text{(chain rule)} \\ &= P(X_1|, Y)P(X_2, Y) && \text{(conditional independence)} \end{aligned}$$

The Naïve Bayes Algorithm

Naïve Bayes is a classifier that assumes conditional independence of the variables X_i given the label Y. For example: $P(X_1|X_2, Y) = P(X_1|Y)$

How does this assumption simplify $P(X_1, X_2|Y)$?

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2, Y) && \text{(chain rule)} \\ &= P(X_1|, Y)P(X_2, Y) && \text{(conditional independence)} \end{aligned}$$

In general:

$$P(X_1, \ldots, X_d|Y) = \prod_i P(X_i|, Y)$$

The Naïve Bayes Algorithm

Naïve Bayes is a classifier that assumes conditional independence of the variables X_i given the label Y. For example: $P(X_1|X_2, Y) = P(X_1|Y)$

How does this assumption simplify $P(X_1, X_2|Y)$?

$$
\begin{aligned}
P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2, Y) && \text{(chain rule)} \\
&= P(X_1|, Y)P(X_2, Y) && \text{(conditional independence)}
\end{aligned}
$$

In general:

$$P(X_1, \ldots, X_d|Y) = \prod_i P(X_i|, Y)$$

How many parameters do we need to describe $P(X_1 \ldots X_n|Y)$?

- Without the conditional independence assumption: $2(2^d - 1)$ and 1
- With the conditional independence assumption:

The Naïve Bayes Algorithm

Naïve Bayes is a classifier that assumes conditional independence of the variables X_i given the label Y. For example: $P(X_1|X_2, Y) = P(X_1|Y)$

How does this assumption simplify $P(X_1, X_2|Y)$?

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2, Y) && \text{(chain rule)} \\ &= P(X_1|, Y)P(X_2, Y) && \text{(conditional independence)} \end{aligned}$$

In general:

$$P(X_1, \ldots, X_d|Y) = \prod_i P(X_i|, Y)$$

How many parameters do we need to describe $P(X_1 \ldots X_n|Y)$?

- Without the conditional independence assumption: $2(2^d - 1)$ and 1
- With the conditional independence assumption: $2d$ and 1

The Naïve Bayes Algorithm

Naïve Bayes assumes conditional independence of the X_i's:
$P(X_1, \ldots, X_d|Y) = \prod_i P(X_i|, Y)$

(more on this assumption soon!)

Using Bayes rule with that assumption:

$$P(Y = y_k|X_1, \ldots, X_d) = \frac{P(Y = y_k) \prod_i P(X_i|Y = y_k)}{P(X)}$$

- Train the algorithm (estimate $P(X_i|Y = y_k)$ and $P(Y = y_k)$)
- To classify, pick the most probable Y^{new} for a new sample $X^{\text{new}} = (X_1^{\text{new}}, X_2^{\text{new}}, \ldots, X_d^{\text{new}})$ as:

$$Y^{\text{new}} \leftarrow \underset{y_k}{\text{argmax}}\; P(Y = y_k) \prod_i P(X_i^{\text{new}}|Y = y_k)$$

Naïve Bayes - Training and Prediction Phase - Discrete $\boldsymbol{X_i}$

Training:

- Estimate $\pi_k \equiv P(Y = y_k)$
- Estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$
    - θ_{ijk} is computed for each label y_k:
        - For each variable X_i:
            - For each value x_{ij} that X_i can take.
    - Example: if X_1 is binary, $P(X_1 | Y = 0)$ is a bernouilli distribution, where:
        - the probability of $X_1 = 0$ given $Y = 0$ is θ_{100}
        - the probability of $X_1 = 1$ given $Y = 0$ is θ_{110}
        - $\theta_{100} = 1 - \theta_{110}$.

Naïve Bayes - Training and Prediction Phase - Discrete X_i

Training:

- Estimate $\pi_k \equiv P(Y = y_k)$, get $\hat{\pi}_k$
- Estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$, get $\hat{\theta}_{ijk}$
    - θ_{ijk} is estimate for each label y_k:
        - For each variable X_i:
            - For each value x_{ij} that X_i can take.

- Prediction: Classify Y^{new}

$$\begin{aligned} Y^{\text{new}} &= \underset{y_k}{\text{argmax}}\ P(Y = y_k) \prod_i P(X_i^{\text{new}} = x_j^{\text{new}} | Y = y_k) \\ &= \underset{y_k}{\text{argmax}}\ \pi_k \prod_i \theta_{i, X_i^{\text{new}}, k} \end{aligned}$$

But... how do we estimate these parameters?

Naïve Bayes - Training Phase - Discrete X_i - Maximum (Conditional) Likelihood Estimation

$P(X|Y = y_k)$ has parameters θ_{ijk}, one for each value x_{ij} of each X_i.

To follow the MLE principle, we pick the parameters θ that maximizes the **conditional** likelihood of the data given the parameters.

To estimate:

- Compute
$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D(Y = y_k)}{|D|}$$
- For each label y_k:
  - For each variable X_i:
    - For each value x_{ij} that X_i can take, compute:
$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D(X_i = x_{ij} \wedge Y = y_k)}{\#D(Y = y_k)}$$
.

Let’s train!

O Is older than 35 years	I Has Personal Income	S Is a Student	J Birthday before July 1st	Y Buys computer
0	1	0	0	0
0	1	0	1	0
1	1	0	1	1
1	1	0	1	1
1	0	1	0	1
1	0	1	1	0
0	0	1	0	1
0	1	0	0	0
0	0	1	1	1
0	1	1	0	1
0	0	1	1	1
1	1	0	1	1
0	1	1	0	1
1	1	0	0	0

Let’s train!

Removed one variable to simplify the problem + changed order of samples.

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

O	*S*	*J*	*Y*
0	0	0	**0**
0	0	1	**0**
0	0	0	**0**
1	1	1	**0**
1	0	0	**0**
0	1	0	**1**
1	0	1	**1**
1	0	1	**1**
1	1	0	**1**
0	1	1	**1**
0	1	0	**1**
0	1	1	**1**
1	0	1	**1**
0	1	0	**1**

Let’s train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$\boldsymbol{Y = 1}$		$\boldsymbol{Y = 0}$	
$P(Y = 1) =$		$P(Y = 0) =$	
$$P(O=1	Y=1) = $$	$$P(O=1	Y=0) = $$
$$P(O=0	Y=1) = $$	$$P(O=0	Y=0) = $$
$$P(S=1	Y=1) = $$	$$P(S=1	Y=0) = $$
$$P(S=0	Y=1) = $$	$$P(S=0	Y=0) = $$
$$P(J=1	Y=1) = $$	$$P(J=1	Y=0) = $$
$$P(J=0	Y=1) = $$	$$P(J=0	Y=0) = $$

Let’s train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$		$Y = 0$	
$P(Y = 1) = 9/14$		$P(Y = 0) = 5/14$	
$$P(O=1	Y=1) = $$	$$P(O=1	Y=0) = $$
$$P(O=0	Y=1) = $$	$$P(O=0	Y=0) = $$
$$P(S=1	Y=1) = $$	$$P(S=1	Y=0) = $$
$$P(S=0	Y=1) = $$	$$P(S=0	Y=0) = $$
$$P(J=1	Y=1) = $$	$$P(J=1	Y=0) = $$
$$P(J=0	Y=1) = $$	$$P(J=0	Y=0) = $$

Let’s train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$	$Y = 0$		
P(Y=1) = 9/14	P(Y=0) = 5/14		
P(O=1\	Y=1) = 4/9</font>	P(O=1\	Y=0) =
P(O=0\	Y=1) =	P(O=0\	Y=0) =
P(S=1\	Y=1) =	P(S=1\	Y=0) =
P(S=0\	Y=1) =	P(S=0\	Y=0) =
P(J=1\	Y=1) =	P(J=1\	Y=0) =
P(J=0\	Y=1) =	P(J=0\	Y=0) =

Let’s train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$		$Y = 0$	
P(Y=1) = 9/14		P(Y=0) = 5/14	
P(O=1\	Y=1) = 4/9</font>	P(O=1\	Y=0) = 2/5 </font>
P(O=0\	Y=1) = 5/9 </font>	P(O=0\	Y=0) =
P(S=1\	Y=1) =	P(S=1\	Y=0) =
P(S=0\	Y=1) =	P(S=0\	Y=0) =
P(J=1\	Y=1) =	P(J=1\	Y=0) =
P(J=0\	Y=1) =	P(J=0\	Y=0) =

Let’s train!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$		$Y = 0$		
P(Y=1) = 9/14		P(Y=0) = 5/14		
P(O=1\	Y=1) = 4/9</font>		P(O=1\	Y=0) = 2/5 </font>
P(O=0\	Y=1) = 5/9 </font>		P(O=0\	Y=0) = 3/5
P(S=1\	Y=1) = 6/9		P(S=1\	Y=0) = 1/5
P(S=0\	Y=1) = 3/9		P(S=0\	Y=0) = 4/5
P(J=1\	Y=1) = 5/9		P(J=1\	Y=0) = 2/5
P(J=0\	Y=1) = 4/9		P(J=0\	Y=0) = 3/5

Let’s predict!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$		$Y = 0$		
P(Y=1) = 9/14		P(Y=0) = 5/14		
P(O=1\	Y=1) = 4/9</font>		P(O=1\	Y=0) = 2/5 </font>
P(O=0\	Y=1) = 5/9 </font>		P(O=0\	Y=0) = 3/5
P(S=1\	Y=1) = 6/9		P(S=1\	Y=0) = 1/5
P(S=0\	Y=1) = 3/9		P(S=0\	Y=0) = 4/5
P(J=1\	Y=1) = 5/9		P(J=1\	Y=0) = 2/5
P(J=0\	Y=1) = 4/9		P(J=0\	Y=0) = 3/5

Let’s predict!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$	$Y = 0$		
P(Y=1) = 9/14	P(Y=0) = 5/14		
P(O=1\	Y=1) = 4/9</font>	P(O=1\	Y=0) = 2/5 </font>
P(O=0\	Y=1) = 5/9 </font>	P(O=0\	Y=0) = 3/5
P(S=1\	Y=1) = 6/9	P(S=1\	Y=0) = 1/5
P(S=0\	Y=1) = 3/9	P(S=0\	Y=0) = 4/5
P(J=1\	Y=1) = 5/9	P(J=1\	Y=0) = 2/5
P(J=0\	Y=1) = 4/9	P(J=0\	Y=0) = 3/5

Let's predict!

O= Is older than 35, S= Is a student, J = Birthday before July 1, Y= buys computer

$Y = 1$		$Y = 0$	
P(Y=1) = 9/14		P(Y=0) = 5/14	
P(O=1\	Y=1) = 4/9</font>	P(O=1\	Y=0) = 2/5 </font>
P(O=0\	Y=1) = 5/9 </font>	P(O=0\	Y=0) = 3/5
P(S=1\	Y=1) = 6/9	P(S=1\	Y=0) = 1/5
P(S=0\	Y=1) = 3/9	P(S=0\	Y=0) = 4/5
P(J=1\	Y=1) = 5/9	P(J=1\	Y=0) = 2/5
P(J=0\	Y=1) = 4/9	P(J=0\	Y=0) = 3/5

$$Y^{\text{new}} = \underset{y_k}{\text{argmax}}\; P(Y = y_k) P(O = 0, S = 1, J = 1 | Y = y_k)$$

$$= \underset{y_k}{\text{argmax}}\; P(Y = y_k) P(O = 0 | Y = y_k) P(S = 1 | Y = y_k) P(J = 1 | Y = y_k)$$

P(Y=1) P(O=0| Y=1)P(S = 1 | Y=1) P(J = 1 | Y=1) = 9/14 * 5/9 * 6/9 * 5/9 =0.132

P(Y=0) P(O=0| Y=0)P(S = 1 | Y=0) P(J = 1 | Y=0) = 5/14 * 3/5 * 1/5 * 2/5 = 0.017

Pick label 1!

Can also compute P(Y|X):

Not required to make predictions, but we have everything we need:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(X) = \sum_k P(X|Y = y_k)P(Y = y_k)$$

P(Y=1) P(O=0| Y=1)P(S = 1 | Y=1) P(J = 1 | Y=1) = 0.132

P(Y=0) P(O=0| Y=0)P(S = 1 | Y=0) P(J = 1 | Y=0) = 0.017

P(Y=1 \| O=0, S = 1, J = 1) = 0.886

P(Y=0 \| O=0, S = 1, J = 1) = 0.114

Classification Accuracy

- To get an estimate of generalization performance, compute accuracy on held-out set (more about this soon).
    - Never train on your test data!
- Assume you train and use this algorithm to predict "Buy Computer?" with a large dataset, and obtain binary classification accuracy of 75%.
    - What does this mean? Is it good or bad?

Classification Accuracy

- What if 70% is the probability of $Y = 1$. Is 75% impressive?
    - What is an easy way to obtain 70\% classification accuracy?

Classification Accuracy

- What if 70% is the probability of $Y = 1$. Is 75% impressive?
    - What is an easy way to obtain 70\% classification accuracy?
- What is chance performance?

Classification Accuracy

- What if 70% is the probability of $Y = 1$. Is 75% impressive?
    - What is an easy way to obtain 70\% classification accuracy?
- What is chance performance?
- What is the accuracy if I flip an unbiased coin?

Classification Accuracy

- What if 70% is the probability of $Y = 1$. Is 75% impressive?
    - What is an easy way to obtain 70\% classification accuracy?
- What is chance performance?
- What is the accuracy if I flip an unbiased coin?
- If P(Y=1)>0.5, then we can just predict 1 all the time!
    - What will be the accuracy?

Classification Accuracy

- What if 70% is the probability of $Y = 1$. Is 75% impressive?
    - What is an easy way to obtain 70\% classification accuracy?
- What is chance performance?
- What is the accuracy if I flip an unbiased coin?
- If P(Y=1)>0.5, then we can just predict 1 all the time!
    - What will be the accuracy?
- What happens if you predict Y=1 with probability 0.7 ==> this is called probability matching in cognitive science

Naïve Bayes observation 1

Usually the X_i are not conditionally independent:

$$P(X_1, \ldots, X_d|Y) \neq \prod_i P(X_i|, Y)$$

- Even if the "naïve" conditional independence assumption is not true in the data, Naïve Bayes might still perform well and is used anyways
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- To see the effect of the violation of this assumption, consider the extreme case in which X_i is a copy of X_k. What is the effect on $P(Y|X)$?

Naïve Bayes observation 1

Usually the X_i are not conditionally independent:

$$P(X_1, \ldots, X_d|Y) \neq \prod_i P(X_i|, Y)$$

- Even if the "naïve" conditional independence assumption is not true in the data, Naïve Bayes might still perform well and is used anyways
    - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- To see the effect of the violation of the conditional independence assumption, consider the extreme case in which X_i is a copy of X_k. What is the effect on $P(Y|X)$?

$$\frac{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1)}{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1) + P(Y=0)\ P(O=0|Y=0)\ P(S=1|Y=0)\ P(J=1|Y=0)}$$

- To see the effect of the violation of the conditional independence assumption, consider the extreme case in which X_i is a copy of X_k. What is the effect on $P(Y|X)$?

$$\frac{P(Y=1)\ P(O=0|Y=1)\ P(O'=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1)}{P(Y=1)\ P(O=0|Y=1)\ P(O'=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1) + \ldots \\ \ldots\ P(Y=0)\ P(O=0|Y=0)\ P(O'=0|Y=0)\ P(S=1|Y=0)\ P(J=1|Y=0)}$$

Consider for example that $P(O=1|Y=1) > P(O=1|Y=0)$, how does $P(Y=1|O=1, O'=1, S=1, J=1)$ with the duplicated variable compare to respect $P(Y=1|O=1, S=1, J=1)$?

Naïve Bayes observation 2

What if we have an irrelevant variable?

$$\frac{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1)}{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1) + P(Y=0)\ P(O=0|Y=0)\ P(S=1|Y=0)\ P(J=1|Y=0)}$$

Naïve Bayes observation 2

What if we have an irrelevant variable?

$$\frac{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1)}{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1) + P(Y=0)\ P(O=0|Y=0)\ P(S=1|Y=0)\ P(J=1|Y=0)}$$

Assume J is independent of Y: $P(J|Y = 0) = P(J|Y = 1) = P(J)$

Does it hurt classification?

Naïve Bayes observation 2

What if we have an irrelevant variable?

$$\frac{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1)}{P(Y=1)\ P(O=0|Y=1)\ P(S=1|Y=1)\ P(J=1|Y=1) + P(Y=0)\ P(O=0|Y=0)\ P(S=1|Y=0)\ P(J=1|Y=0)}$$

Assume J is independent of Y: $P(J|Y=0) = P(J|Y=1) = P(J)$

Does it hurt classification?

- If we have the correct estimates, then performance is not affected.
- If we have noisy estimates, performance is affected.</font>

Naïve Bayes observation 3

Another way to view Naïve Bayes with Boolean Y and X_is is:

Decision rule: is this quantity greater or less than 1?

$$\frac{P(Y=1|X_1\ldots X_d)}{P(Y=0|X_1\ldots X_d)} = \frac{P(Y=1)P(X_1\ldots X_d|Y=1)}{P(Y=0)P(X_1\ldots X_d|Y=0)} \qquad > \text{ or } < 1?$$

Practical concern: What happens when d is large?

Naïve Bayes observation 3

Another way to view Naïve Bayes with Boolean Y and X_is is:

Decision rule: is this quantity greater or less than 1?

$$\frac{P(Y=1|X_1\ldots X_d)}{P(Y=0|X_1\ldots X_d)} = \frac{P(Y=1)P(X_1\ldots X_d|Y=1)}{P(Y=0)P(X_1\ldots X_d|Y=0)} \qquad > \text{ or } < 1?$$

Taking the log of this ratio prevents **underflow** and expresses the decision rule in a useful way (we will see later more reasons why it's useful).

$$\ln\frac{P(Y=1|X_1\ldots X_d)}{P(Y=0|X_1\ldots X_d)} = \ln\frac{P(Y=1)}{P(Y=0)} + \sum_i \ln\frac{P(X_i|Y=1)}{P(X_i|Y=0)} \qquad > \text{ or } < 0?$$

Since X_is are boolean, we can simplify the notation:

- $\theta_{ik} = \hat{P}(X_i = 1|Y=k)$
- $1-\theta_{ik} = \hat{P}(X_i = 0|Y=k)$

$$\begin{aligned}\ln\frac{P(Y=1|X_1\ldots X_d)}{P(Y=0|X_1\ldots X_d)} &= \ln\frac{P(Y=1)}{P(Y=0)} + \sum_i \ln\frac{P(X_i|Y=1)}{P(X_i|Y=0)} \\ &= \ln\frac{P(Y=1)}{P(Y=0)} + \sum_i X_i \ln\frac{\theta_{i1}}{\theta_{i0}} + \sum_i (1-X_i)\ln\frac{1-\theta_{i1}}{1-\theta_{i0}}\end{aligned}$$

Naïve Bayes observation 4

If unlucky, our MLE estimate for $P(X_i|Y)$ might be zero.

- for example, X_i = birthdate. xi = Jan_25_1992.

- Why worry about just one parameter out of many?

Naïve Bayes observation 4

If unlucky, our MLE estimate for $P(X_i|Y = y_k)$ might be zero.

- for example, X_i = birthdate. xi = Jan_25_1992.

- Why worry about just one parameter out of many?

$$P(Y = y_k)P(X_1|Y = y_k)P(X_2|Y = y_k)\ldots P(X_d|Y = y_k)$$

```
What happens if one of the $P(X_i|Y=y_k)$ is zero?
```

- What can be done to address this?

Naïve Bayes - Training Phase - Discrete $\boldsymbol{X_i}$

Method 1: Maximum (Conditional) Likelihood Estimation

$P(X|Y = y_k)$ has parameters θ_{ijk}, one for each value x_{ij} of each X_i.

To follow the MLE principle, we pick the parameters θ that maximizes the **conditional** likelihood of the data given the parameters.

Method 2: Maximum A Posteriori Probability Estimation

To follow the MAP principle, pick the parameters θ with maximum posterior probability given the conditional likelihood of the data and the prior on θ.

Naïve Bayes - Training Phase - Discrete $\boldsymbol{X_i}$

Method 1: Maximum (Conditional) Likelihood Estimation

To estimate:

- Compute

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D(Y = y_k)}{|D|}$$

- For each label y_k:
    - For each variable X_i:
        - For each value x_{ij} that X_i can take, compute:

          $$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D(X_i = x_{ij} \wedge Y = y_k)}{\#D(Y = y_k)}$$

          .

Method 2: Maximum A Posteriori Probability Estimation (Beta or Dirichlet priors)

- K: the number of values Y can take
- J: the number of values X can take (we assume here that all X_j have the same number of possible values, but this can be changed)

- Example prior for π_k where K > 2:
  - $\mathrm{Dirichlet}(\beta_\pi, \beta_\pi, \ldots, \beta_\pi)$ prior. (optionally, you can choose different values for each parameter to encode a different weighting).
  - if $K = 2$ this becomes a Beta prior.

- Example prior for θ_{ijk} where J>2 :
  - $\mathrm{Dirichlet}(\beta_\theta, \beta_\theta, \ldots, \beta_\theta)$ prior. (optionally, you can choose different values for each parameter to encode a different weighting, you can choose a different prior per X_i or even per label y_k).
  - if $J = 2$ this becomes a Beta prior.

Method 2: Maximum A Posteriori Probability Estimation (Beta or Dirichlet priors)

- K: the number of values Y can take
- J: the number of values X can take

These priors will act as imaginary examples that smooth the estimated distributions and prevent zero values.

To estimate:

- Compute

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D(Y = y_k) + (\beta_\pi - 1)}{|D| + K(\beta_\pi - 1)}$$

- For each label y_k:
  - For each variable X_i:
    - For each value x_{ij} that X_i can take, compute:

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D(X_i = x_{ij} \wedge Y = y_k) + (\beta_\theta - 1)}{\#D(Y = y_k) + J(\beta_\theta - 1)}$$

.