Discrete Probability Review and Naive Bayes

Rob McCulloch

1. Probability and Machine Learning
2. Discrete Random Variables
3. Conditional, Joint and Marginal Distributions
4. Bayes Theorem
5. Several Variables
6. Independence
7. IID
8. Bayes Rule Classification
9. Naive Bayes Classification
10. Sentiment Analysis: Spam or Ham

1. Probability and Machine Learning

In Machine Learning we often divide our tasks into *directed* or *undirected* problems:

In directed problems we have variables x and y and we want to predict y from x.

We can do this using ideas from probability:

- specify $p(y \mid x)$, the conditional distribution of y given x. (e.g. logistic regression, linear regression with normal errors).
- classification with Bayes theorem: $p(y)$, $p(x \mid y) \Rightarrow p(x, y) \Rightarrow p(y \mid x)$. (e.g Naive Bayes, LDA: linear discriminant analysis).

This seems natural in that just knowing x typically does leave us sure about y, that is we do not have a deterministic model $y = f(x)$.

Note:

When y is a discrete variable (e.g binary) we have a classification problem.

E.g. y =(is a text message spam or not) given x = (the text of the message).

Note:

You don't have to use probability models to do Machine Learning.

You can just cook up an algorithm and see how well it works (e.g K nearest neighbors).

Note:

In classification problems,

the $p(y \mid x)$ approach is often called *discriminative model*, while the Bayes theorem approach with $p(x, y)$ is often called a *generative model*.

But, the terminology is a little confusing in the literature. Sometimes a "generative model" just means you have a probability model.

In undirected problems we just have x but we look for simplifying structure in x.

- clustering, look for groups of similar observations.
- estimate $p(x)$ (mixture of normals).
- LDA: latent dirichlet allocation (very popular model for text data).

In these notes we will briefly review very basic discrete probability concepts and then learn the Naive Bayes algorithm, which uses the Bayes theorem approach.

Don't confuse Bayes Theorem with *Bayesian statistics*, that is a different thing.

2. Discrete Random Variables

A random variable is *a number we're not sure about*.

Its *distribution* describes what we think it might turn out to be.

For a discrete random variable, we specify the distribution by:

- Listing all the possible numbers it can turn out to be.
- Assigning a probability to each possible outcome.
- Each probability is between 0 and 1.
- The probabilities add up to 1.

Note: "discrete" refers to the situation where can make the list (we have a countable set of possible outcomes).
Later we will look at continuous random variable where such a list is not practical.

Example:

Suppose we are about to toss two coins.

Let X denote the number of heads.

Then the distribution of X might be given by

x	$P(X = x)$
0	.25
1	.5
2	.25

Notation: $P(X = x)$ is "the probability X turns out to be x".
You will see other notations !!

Example:

Let S denote sales next period (thousands of units).

Then the distribution of S might be given by

s	$P(S = s)$
1	.095
2	.23
3	.44
4	.235

What is $P(S > 2)$?
What is $P(S \geq 2)$?

The Bernoulli Distribution

A very common situation is that we are wondering whether something will happen or not.

Heads or tails, respond or don't respond, ...

It turns out to be very convenient to code up one possibility as a 0, and the other possibility as a 1.

The gives us the *Bernoulli distribution*.

$X \sim$ Bernoulli(p) means:

x	$P(X = x)$
0	1-p
1	p

Example:

You are about to toss a coin.

Let X be 1 if it comes up Heads and 0 if tails.

$$X \sim \text{Bernoulli}(.5).$$

Example:

You are about to target a customer.

Let R be 1 if the respond (buy) and 0 otherwise.

For a particular customer we might have:

$$R \sim \text{Bernoulli}(.05)$$

3. Conditional, Joint, and Marginal Distributions

In general we want to use probability to address problems involving more than one variable at the time.

For example we may need want to:

- describe our uncertainty about several quantities together
  (the joint distribution)
- understand how learning values about some variables affects our beliefs about others
  (the conditional distribution).

Suppose you are thinking about sales next quarter.
In order to think about sales, it may be helpful to think about sales *and* what will happen to the economy.

Let E denote the performance of the economy next quarter... for simplicity, say $E = 1$ if the economy is expanding and $E = 0$ if the economy is contracting (what kind of random variable is this?)

Let's assume $p(E = 1) = 0.7$

Let S denote my sales next quarter... and let's suppose the following probability statements:

s	$p(S = s\|E = 1)$	s	$p(S = s\|E = 0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

These are called *Conditional Distributions*

s	$p(S = s \mid E = 1)$	s	$p(S = s \mid E = 0)$
1	0.05	1	0.20
2	0.20	2	0.30
3	0.50	3	0.30
4	0.25	4	0.20

- In blue is the conditional distribution of S given $E = 1$
- In red is the conditional distribution of S given $E = 0$
- We read: *the probability of Sales of 4 ($S = 4$)* **given (or conditional on)** *the economy is growing ($E = 1$) is 0.25.*

The way S is related to E is captured by the difference between the conditional distributions !!!

The conditional distributions tell us about about what can happen to S for a given value of E... but what about S and E jointly?

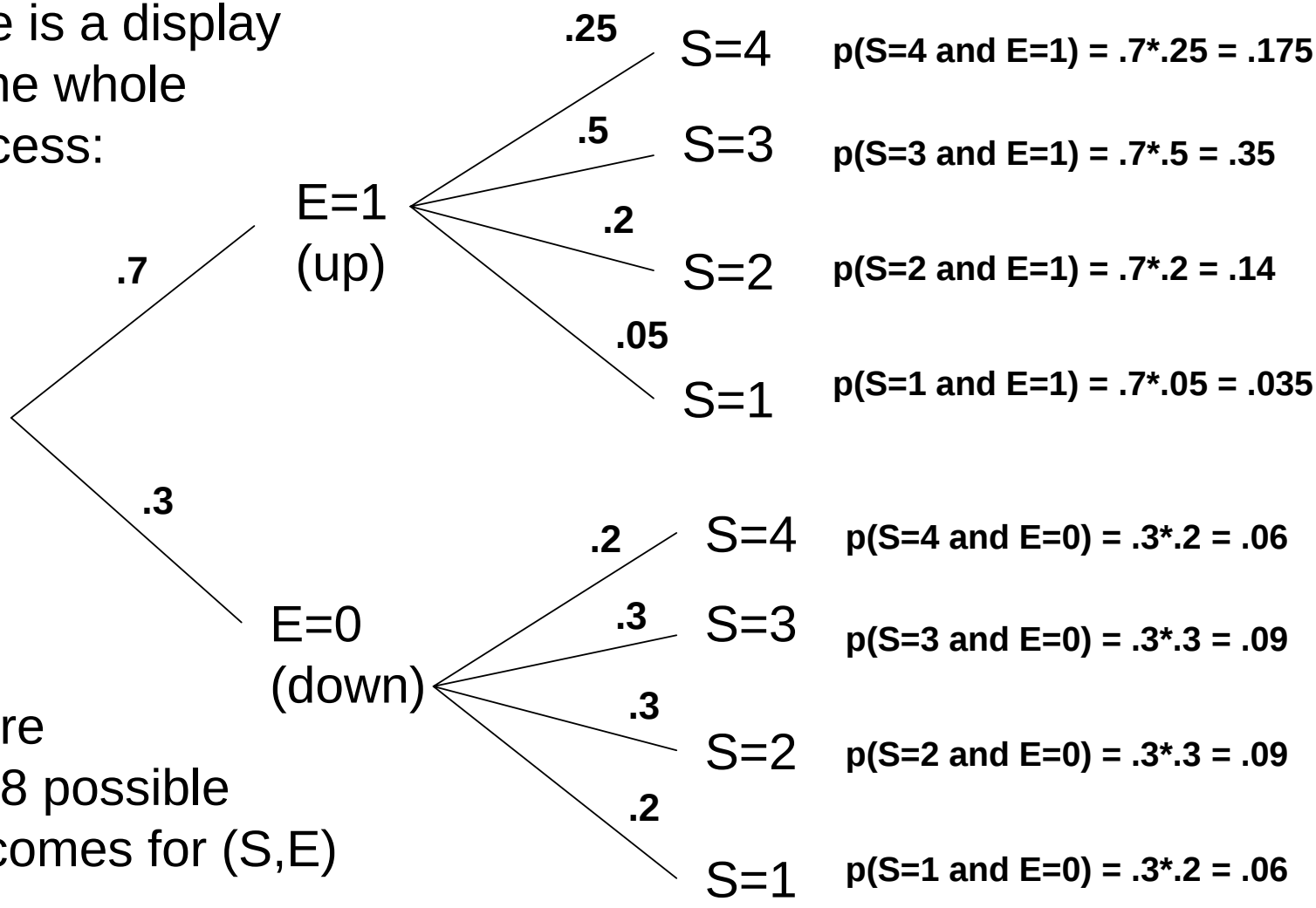
$$\begin{aligned} p(S = 4 \quad \text{and} \quad E = 1) &= p(E = 1) \times p(S = 4|E = 1) \\ &= 0.70 \times 0.25 = 0.175 \end{aligned}$$

In English, 70% of the times the economy grows and 1/4 of those times sales equals 4... 25% of 70% is 17.5.

Notation:

$P(S = 4, E = 1)$ is the same as $p(S = 4 \quad \text{and} \quad E = 1)$.

here is a display of the whole process:



There are 8 possible outcomes for (S,E)

We can specify the distribution of the pair of random variables (S, E) by listing all possible pairs and the corresponding probability.

(s, e)	$p(S = s, E = e)$
(1,1)	.035
(2,1)	.14
(3,1)	.35
(4,1)	.175
(1,0)	.06
(2,0)	.09
(3,0)	.09
(4,0)	.06

Question: What is $Pr(S = 1)$?

We call the probabilities of E and S together the joint distribution of E and S.

In general the notation is...

- $p(Y = y, X = x)$ is the joint probability the random variable Y equals y **AND** the random variable X equals x.
- $p(Y = y|X = x)$ is the conditional probability the random variable Y takes the value y **GIVEN** that X equals x.
- $p(Y = y)$ or $p(X = x)$ are the marginal probabilities of $Y = y$ or $X = x$

Important relationships

Relationship between the joint and conditional...

$$\begin{aligned} Pr(Y = y, X = x) &= Pr(X = x) \times Pr(Y = y|X = x) \\ &= Pr(Y = y) \times Pr(X = x|Y = y) \end{aligned}$$

Relationship between joint and marginal...

$$\begin{aligned} Pr(X = x) &= \sum_y Pr(X = x, Y = y) \\ Pr(Y = y) &= \sum_x Pr(X = x, Y = y) \end{aligned}$$

A Note on Notation

We have used the notations

$$P(Y = y), \quad P(Y = y, X = x), \quad P(Y = y \mid X = x)$$

You will see all kinds of similar, but not exactly the same notations for these fundamental concepts.

For example, we will sometimes use $p(x, y)$ in place of $P(X = x, Y = y)$ when it is clear from the context what we mean.

For example in the (S, E) example I could write $P(S = s, E = e)$ or just $p(s, e)$.

The Two-Way Table Display of the Joint Distribution

This is a nice way to display a joint distribution.
Same information as when we just listed the (s, e) pairs and their probabilities but this way we can see some things more easily.

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

For example, you can see why the marginals are called "the marginals".

Conditionals from Joints

We derived the joint distribution of (E, S) from the marginal for E and the conditional $S \mid E$.

You can also calculate the conditional from the joint by doing it the other way

$$Pr(Y = y, X = x) = Pr(X = x)\, Pr(Y = y \mid X = x)$$

$\Rightarrow$

$$Pr(Y = y \mid X = x) = \frac{Pr(Y = y, X = x)}{Pr(X = x)}$$

Example... Given $E = 1$ what is the probability of $S = 4$?

		S				
		1	2	3	4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$p(S = 4|E = 1) = \frac{p(S = 4, E = 1)}{p(E = 1)} = \frac{0.175}{0.7} = 0.25$$

Example... Given $S = 4$ what is the probability of $E = 1$?

		S = 1	S = 2	S = 3	S = 4	
E	0	.06	.09	.09	.06	.3
	1	.035	.14	.35	.175	.7
		.095	.23	.44	.235	1

$$p(E = 1|S = 4) = \frac{p(S = 4, E = 1)}{p(S = 4)} = \frac{0.175}{0.235} = 0.745$$

4. Bayes Theorem

So, in general, you can compute the joint from marginals and conditionals and the other way around.

How you think about stuff depends on what's easiest or what you know, or what you care about.

Suppose you toss two coins: X is the first, Y is the second.
In each case 1 means a head and 0 a tail.
What is $P(X = 1, Y = 1) =$ P(two heads) ?

(1) Directly figure out the joint distribution.
There are 4 possible outcomes for the two coins and each is equally likely so it is $1/4$.

(2) Figure out some marginals and conditionals.
$P(X = 1, Y = 1) = P(X = 1) * P(Y = 1 \mid X = 1) =$
$(1/2) * (1/2) = 1/4.$

Bayes Theorem refers to the situation where we build a model for (Y, X) by thinking about

$$Pr(Y = y \mid X = x), \quad Pr(X = x).$$

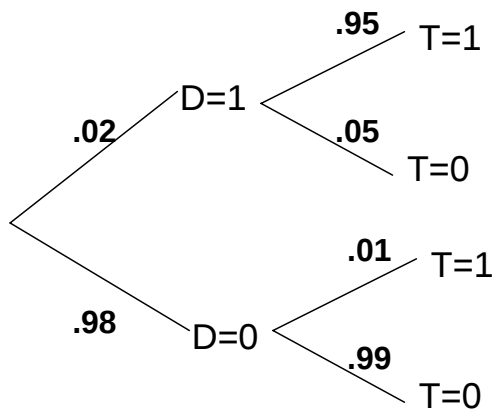
and then, having observed $Y = y$ compute

$$Pr(X = x \mid Y = y)$$

Example: Disease Testing

Disease testing example...

Let $D = 1$ indicate you have a disease.
Let $T = 1$ indicate that you test positive for it
We know the marginal of D and the conditional of T given D.



$\Longrightarrow$

		D = 0	D = 1
T	0	.9702=.98*.99	.001
	1	.0098	.019

If you take the test and the result is positive, you are really interested in the question: Given that you tested positive, what is the chance you have the disease?

		D	
		0	1
T	0	.9702	.001
	1	.0098	.019

$$p(D = 1|T = 1) = \frac{0.019}{(0.019 + 0.0098)} = 0.66$$

The computation of $p(y|x)$ from $p(y)$ and $p(x|y)$ is called Bayes theorem...

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\sum_y p(x,y)} = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

In the disease testing example:

$$p(D=1|T=1) = \frac{p(T{=}1|D{=}1)p(D{=}1)}{p(T{=}1|D{=}1)p(D{=}1)+p(T{=}1|D{=}0)p(D{=}0)}$$

$$p(D=1|T=1) = \frac{0.019}{(0.019+0.0098)} = 0.66$$

Note:

A nice way to think about Bayes theorem is with the *odds ratio*:

$$p(y \mid x) = \frac{p(x, y)}{p(x)} \propto p(y)\, p(x \mid y)$$

In the y binary case we have

$$\frac{p(Y = 1 \mid x)}{p(Y = 0 \mid x)} = \frac{p(Y = 1)}{p(Y = 0)} \frac{p(x \mid Y = 1)}{p(x \mid Y = 0)}$$

The posterior odds ratio =
the prior odds ratio $\times$ the likelihood ratio.

$\frac{p(Y=1)}{p(Y=0)}$: the prior odds ratio.

$\frac{p(x|Y=1)}{p(x|Y=0)}$: the likelihood ratio.

$\frac{p(Y=1|x)}{p(Y=0|x)}$: the posterior odds ratio.

Disease Testing:

$$\frac{p(Y=1|x)}{p(Y=0|x)} = \frac{p(Y=1)}{p(Y=0)} \frac{p(x|Y=1)}{p(x|Y=0)}$$

$\frac{p(Y=1)}{p(Y=0)}$: the prior odds ratio.

$\frac{p(x|Y=1)}{p(x|Y=0)}$: the likelihood ratio.

$\frac{p(Y=1|x)}{p(Y=0|x)}$: the posterior odds ratio.

Disease testing:
posterior odds: .66/(1-.66) = 1.941176
prior odds: .02/.98 = 0.02040816
likelihood ratio: .95/.01 = 95

prior odds x likelihood ratio: .0204*95 = 1.938

Probability from odds:

$$p(Y=1|x) = \frac{p(Y=1)\,p(x|Y=1)}{p(Y=0)\,p(x|Y=0) + p(Y=1)\,p(x|Y=1)}$$

Divide top and bottom by $p(Y=0)\,p(x|Y=0)$:

$$p(Y=1|x) = \frac{odds}{1+odds}$$

Disease testing:
$1.938/(1+1.938) = 0.6596324$

5. Several Variables

Of course, we often want to think about more than two variables at a time.

We can extend the ideas we used with two variables to many variables.

Suppose we have the three random variables,

$$(Y_1, Y_2, Y_3)$$

Then,

$$Pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) =$$

$$Pr(Y_1 = y_1)\, Pr(Y_2 = y_2 \mid Y_1 = y_1)\, Pr(Y_3 = y_3 \mid Y_1 = y_1, Y_2 = y_2)$$

Or, using more succinct notation:

$$p(y_1, y_2, y_3) = p(y_1)\, p(y_2 \mid y_1)\, p(y_3 \mid y_1, y_2)$$

You can keep going for any number of variables !!

Example

Suppose we have 10 voters.
4 are republican and 6 are democratic.

We randomly choose 3.

Let Y_i be 1 if the i^{th} voter is a democrat and 0 otherwise,
$i = 1, 2, 3$.

What is

$$Pr(Y_1 = 1, Y_2 = 1, Y_3 = 1)$$

What is the probability of getting three democrats in a row ??

$$Pr(Y_1 = 1, Y_2 = 1, Y_3 = 1) =$$

$$Pr(Y_1 = 1)\, p(Y_2 = 1 \mid Y_1 = 1)\, Pr(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$$

$$= (6/10)(5/9)(4/8)$$

$$= (1/6) = .167.$$

When we randomly pick a person from a population of people, and then randomly pick a second from the ones left, and so on, we call it *sampling without replacement*.

If we put the person back each time and randomly choose from the whole group each time, then we call it *sampling with replacement*.

Suppose we sample from our 10 voters with replacement.

Now what is

$$Pr(Y_1 = 1, Y_2 = 1, Y_3 = 1)$$

$$Pr(Y_1 = 1, Y_2 = 1, Y_3 = 1) =$$

$$Pr(Y_1 = 1)\, p(Y_2 = 1 \mid Y_1 = 1)\, Pr(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$$

$$= (6/10)(6/10)(6/10)$$

$$= .6^3 = .216$$

Notice that when we sample with replacement

$$p(Y_2 = 1 \mid Y_1 = y_1) \text{ and } Pr(Y_3 = 1 \mid Y_1 = y_1, Y_2 = y_2)$$

do not depend on y_1 and y_2.

What happens for Y_1 does not affect what we think will happen for Y_2 and what happens for Y_1 *and* Y_2 does not affect what will happen for Y_3.

In this case we say the random variables are *independent*.

6. Independence

Given a bunch of random variables, we say they are independent of each other if the conditional distribution of any one of them does not depend on anything you might observe for any of the others.

Example

Suppose I am about to toss 100 coins.

Let Y_i be 1 if the i^{th} coin is a head and 0 otherwise.

What is $Pr(Y_3 = 1)$?

What is $Pr(Y_3 = 1 \mid Y_1 = 1, Y_2 = 0)$?
What is $Pr(Y_3 = 1 \mid Y_1 = 0, Y_2 = 1)$?
What is $Pr(Y_3 = 1 \mid Y_1 = 0, Y_2 = 0)$?
What is $Pr(Y_3 = 1 \mid Y_1 = 1, Y_2 = 1)$?

What is

$Pr(Y_{100} = 1 \mid Y_1 = 1, Y_2 = 1, \ldots, Y_{99} = 1)$ first 99 are heads?

What is

$Pr(Y_1 = 1, Y_2 = 1, Y_3 = 1, \ldots, Y_{100} = 1)$ 100 heads in a row!!?

Independence, Conditional Equals Marginal

If random variables are independent then the conditional is the marginal.

For two random variables X and Y if X and Y are independent then,

$$Pr(Y = y \mid X = x)$$

does not depend on x.

We also have:

$$Pr(Y = y \mid X = x) = Pr(Y = y)$$

What you believe about Y knowing $X = x$, is the same as what you believe if you know nothing about X.

Example

X is 1 if first coin is head.
Y is 1 if second coin is head.

What is

$P(Y = 1 \mid X = 1)$?

What is

$P(Y = 1 \mid X = 0)$?

What is

$P(Y = 1)$?

Independence, Joints, and Marginals

If X and Y are independent then the joint is the product of the marginals:

$$\begin{aligned} p(x, y) &= p(x)\, p(y \mid x) \\ &= p(x)\, p(y) \end{aligned}$$

This also works "the other way", that is, if the joint is the product of the marginals then they are independent.

Example

You are about to manufacture two parts.

$X = 1$ if part one fails, 0 else.
$Y = 1$ if part two fails, 0 else.

The table below gives the joint distribution of X and Y.

		X	
		0	1
Y	0	.72	.08
	1	.18	.02

		X	
		0	1
Y	0	.72	.08
	1	.18	.02

$Pr(Y = 1 \mid X = 0) = .18/.9 = .2$
$Pr(Y = 1 \mid X = 1) = .02/.1 = .2$
$Pr(Y = 1) = 18 + .02 = .2$

$Pr(Y = 1, X = 1) = .02$
$Pr(Y = 1)Pr(X = 1) = .2 * .1 = .02.$

X and Y are independent.

For random variables Y_i, if they are independent we have

$p(y_1, y_2, \ldots, y_n) =$

$p(y_1)\, p(y_2 \mid y_1)\, p(y_3 \mid y_1, y_2) \ldots p(y_n \mid y_1, y_2, \ldots, y_{n-1})$

$= p(y_1)\, p(y_2)\, p(y_3) \ldots p(y_n)$

Example

If Y_i is 1 if the i^{th} coin is a head, 0 else, $i = 1, 2, \ldots, 10$, what is

$$p(1, 1, \ldots, 1)$$

10 heads in a row?

$p(1, 1, \ldots, 1) =$

$= p(1)\, p(1)\, p(1) \ldots p(1)$

$= .5^{10} = 0.0009765625.$

7. IID

Suppose we are about to toss 100 coins.

Let Y_i be 1 if heads, 0 else,

We usually think the Y's are independent.

In addition, we usually think they are *identically distributed*, that is, each one has *the same marginal distribution*.

What is $Pr(Y_{20} = 1)$?
What is $Pr(Y_{98} = 1)$?

When random variables are independent and identically distributed we say they are **IID**.

I: independent

ID: identically distributed.

In our coins example, each $Y_i \sim \text{Bernoulli}(.5)$.

We can succinctly describe coin tossing by

$$Y_i \sim \text{Bernoulli}(.5), \quad \textit{IID}$$

Example

Suppose we have 10 voters. 4 are republican and 6 are democratic.

We randomly choose 3, sampling *with replacement*.

Let Y_i be 1 if the i^{th} voter is a democrat and 0 otherwise, $i = 1, 2, 3$.

How can we describe the joint distribution of (Y_1, Y_2, Y_3), are they IID?

Example

Suppose we have 10 voters. 4 are republican and 6 are democratic.

We randomly choose 3, sampling *without replacement*.

Let Y_i be 1 if the i th voter is a democrat and 0 otherwise, $i = 1, 2, 3$.

How can we describe the joint distribution of (Y_1, Y_2, Y_3), are they IID?

Example

Suppose we have 1,000,000 voters. 400,000 are republican and 600,000 are democratic.

We randomly choose 3, sampling *without replacement*.

Let Y_i be 1 if the i^{th} voter is a democrat and 0 otherwise, $i = 1, 2, 3$.

How can we describe the joint distribution of (Y_1, Y_2, Y_3), are they IID?

Example

Suppose I am about to toss a die 100 times.

Let D_i be the outcome for the i^{th} toss
(a number in $\{1, 2, 3, 4, 5, 6\}$).

Are the D_i IID?

Example

Suppose an experienced NBA player is about to take repeated free-throws.

Let Y_i be 1 if he makes the i^{th} attempt and 0 otherwise.

Are these Y_i iid Bernoulli?

This is known at the "hot hand" question.

Most people who play sports believe that the can get get "hot" so that if they made the last few, they are more likely to make the next one.

However, if you look at the data, it looks IID Bernoulli!!

How do you look at the data to see if it looks IID Bernoulli.
That is covered in the notes "Modeling with IID Bernoulli Draws".

Example

Suppose the first penalty in an NHL game is on team A.

For subsequent penalties $P_i = 1$ if the penalty is on a different team than the previous one and 0 otherwise.

Are the P's independent?

Are they IID?

These are not IID.

If the last two (or three!) penalties were on the same team, it becomes quite a bit more likely that the next penalty will be on the other team.

See
Reversal of fortune: a statistical analysis of penalty calls in the national hockey league", (2014), Journal of Quantitative Analysis in Sports 10 (2), 207-224 (Jason Abrevaya and Robert McCulloch)

Example

Suppose you are monitoring a stock and for every 10 minute interval, you record whether the price went up or down.

Let U_i be 1 if it goes up in the i^{th} interval, 0 otherwise.

Are the U_i IID?

Of course, this is a much studied question.

We leave this to your finance courses but just note that it is very interesting how little dependence there is!!!!

8. Bayes Rule Classification

Consider the directed data mining problem with a categorical Y.

Many methods can be viewed as an attempt to estimate:

$$p(y|x)$$

the conditional distribution of Y given $X = x$.

For example in logistic regression we have:

$$p(y = 1|x) \sim \text{Bernoulli}(p(x)),\ p(x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}.$$

where $x = (x_1, x_2, \ldots, x_p)'$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$.

An alternative approach is to estimate the full joint distribution of (X, Y) by estimating the marginal for Y ($p(y)$) and the conditional for X ($p(x|y)$).

We then have the joint via:

$$p(x, y) = p(y)\, p(x|y).$$

And classification is then obtained from Bayes Theorem:

$$\begin{aligned} p(y|x) &= \frac{p(y)p(x|y)}{p(x)} \\ &\propto p(y)p(x|y) \end{aligned}$$

As usual we can predict the most probable y or make a decision based on the probabilities.

In general, it seems more straighforward to estimate $p(y|x)$ directly but some well known approaches take the Bayes-rule path.

For example, for x numeric, classic discriminant analysis assumes

$$p(x|y) \sim N(\mu_y, \Sigma).$$

Remember, Y is discrete and we have to estimate $p(x|y)$ for each possible y.

9. Naive Bayes Classification

Naive Bayes classification uses the Bayes Theorem approach to classification.

The tricky part is that we would like this to work for large x!!!

$$x = (x_1, x_2, \ldots, x_p)$$

where p may be large !!!!

In our application we will have $p = 1,136$!!

How do we get $p(x|y)$ from the data when p is large???

Naive Bayes classification simplifies the problem by assumming that the elements of $X = (X_1, X_2, \ldots, X_p)$ are *conditionally independent* given Y:

$$p(x, y) = p(y)\, p(x \mid y) = p(y) \prod_i p(x_i|y)$$

Each coordinate x_i of x gets to multiply in it's own contribution of evidence about y depending on how likely x_i would be if $Y = y$.

For example, suppose we just have $x = (x_1, x_2)$ and each x is binary (0 or 1).

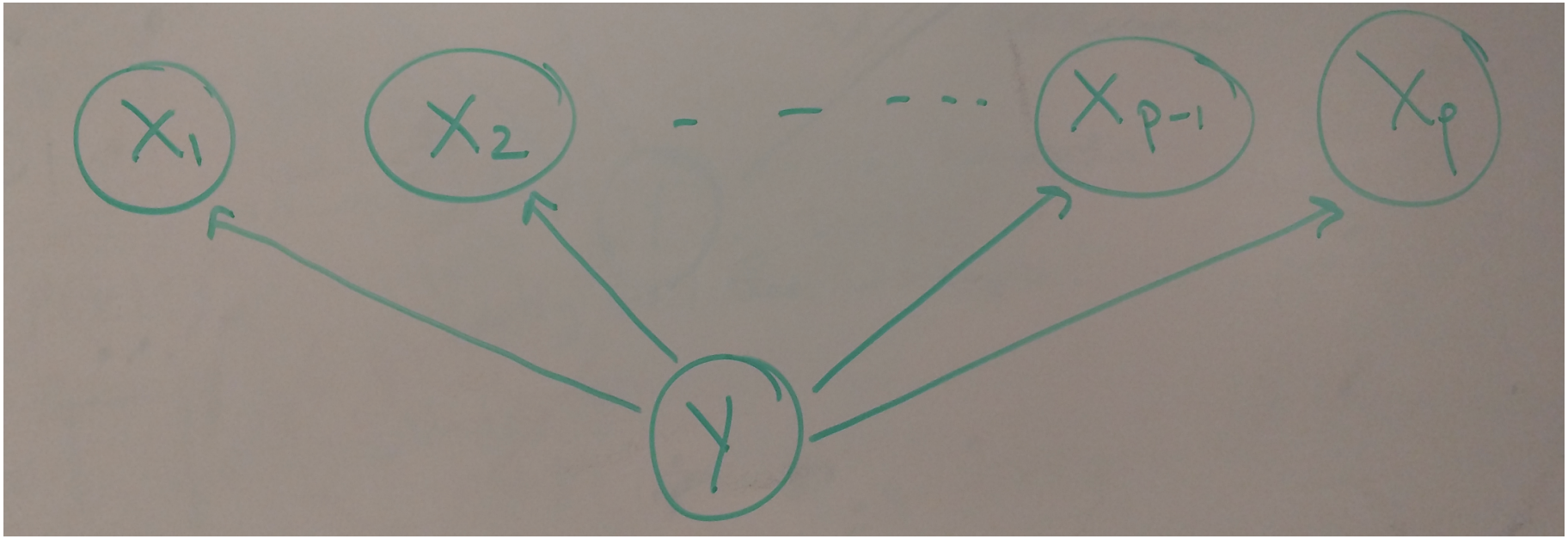
$$p(Y = 1|X_1 = 1, X_2 = 0) =$$

$$= \frac{p(Y = 1)p(X_1 = 1, X_2 = 0|Y = 1)}{p(Y = 1)p(X_1 = 1, X_2 = 0|Y = 1) + p(Y = 0)p(X_1 = 1, X_2 = 0|Y = 0)}$$

$$= \frac{p(Y = 1)p(X_1 = 1|Y = 1)p(X_2 = 0|Y = 1)}{p(Y = 1)p(X_1 = 1|Y = 1)p(X_2 = 0|Y = 1) + p(Y = 0)p(X_1 = 1|Y = 0)p(X_2 = 0|Y = 0)}$$

Same idea works with p x variables instead of 2 !!!
You just have to estimate $p(X_i = 1|y)$ for each i!!!

To draw (X, Y), draw Y and then each $X_i \mid Y$ independently.



Note, this graphical representation of the structure of the joint distribution is a simple but important example of a *graphical model*, more specifically, a directed acyclic graph (DAG). Hopefully we'll get some time to talk more about this towards the end of the course.

NB has some key advantages:

- We only have to estimate the low dimension $p(x_i|y)$ instead of the high dimensional $p(x|y)$!!!!
- Many small bits of information from each x_i can be combined.
- It is simple.

The main disadvantage is that the conditional independence assumption often seems inappropriate. However, *this does not seem to keep from working very well in practice !!!*

According to Mladen Kolar,

> *NB is the single most used classifier out there. NB often performs well, even when the assumption is violated.*

10. Sentiment Analysis: Spam or Ham

Sentiment analysis tries to understand text documents.

A popular approach is to combine "bag of words" with NB.

Each word in the document provides an additional independent piece of evidence about the kind of document it is.

A simple example is trying to classify the document as spam or not: "spam or ham".

Bag of words means just that, we ignore the order of the words.

The document:

> *When the lecture is over, remember to wake up the person sitting next to you in the lecture room.*

is the same as the document,

> *in is lecture lecture next over person remember room sitting the the the to to up wake when you*

SMS Spam Data:

Note: this follows Chapter 4 of "Machine Learning with R", by Brett Lanz.

Note: sms: short message service.

Have 5,559 sms text message documents.

Each one is labelled as spam or ham.

Here is the first (ham) and fourth (spam) observation:

```
> smsRaw[1,]
  type                                               text
1  ham Hope you are having a good week. Just checking in
> smsRaw[4,]
  type
4 spam

4 complimentary 4 STAR Ibiza Holiday or 10,000 cash needs your URGENT collection.
09066364349 NOW from Landline not to lose out! Box434SK38WP150PPM18+
```

Work flow:

- clean: tolower, kill numbers, punctuation, stopwords
- stem: (help,helped,helping,helps) becomes (help,help,help,help)
- tokenization: split a document up into single words (or "tokens" or "terms").
- document term matrix (DTM): rows indicate documents columns are counts for terms.
- train/test split.
- throw away low count terms.
- convert DTM to indicators: Yes if the word (term) is in the document, 0 else.
- do Naive Bayes!!

Note:

Most of the work is processing the data !!!!!

This is typically the case in real world applications.

Getting the data into a form that allows you to analysize it is time consuming and **very** important.

Garbage in, garbage out !!!

In class we will typically focus on getting a basic understanding of the methods and don't emphasize the "data wrangling".

Clean and Stem

Here are the first two documents:

```
> smsRaw$text[1]
[1] "Hope you are having a good week. Just checking in"
> smsRaw$text[2]
[1] "K..give back my thanks."
```

Here are the first 2 docs after cleaning.
smsCC is the cleaned data in the Corpus data structure from the `tm` R package. smsCC is for sms data as a Cleaned Corpus.

```
> smsCC[[1]][1]
$content
[1] "hope good week just check"

> smsCC[[2]][1]
$content
[1] "kgive back thank"
```

did not work too well!!

Tokenize and get DTM

Tokenization gives us 6518 words (or terms) from all the 5,559 sms documents.

The i^{th} row of the DTM gives us the count for each term in document i.

```
> print(dim(smsDtm))
[1] 5559 6518
> library(slam) #for col_sums
> summary(col_sums(smsDtm)) #summarize total time a term is used.
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   6.776   4.000 658.000
> terms = smsDtm$dimnames$Terms
> nterm = length(terms)
> set.seed(14)
> ii = sample(1:nterm,20)
> terms[ii]
 [1] "effect"       "pinku"        "wikipediacom" "mundh"        "wwwsmsconet"
 [6] "marsm"        "voic"         "itz"          "logo"         "hip"
[11] "transfr"      "colin"        "leo"          "technolog"    "text"
[16] "scratch"      "graze"        "prolli"       "tech"         "ofsi"
```

Word frequencies for ham:

Word frequencies for spam:

Train/Test

We split our data into train/test:
train: we estimate/learn/train our model using the training data.
test: see how well we predict on the test data.

```
#train and test
# creating training and test datasets
smsTrain = smsDtm[1:4169, ]
smsTest  = smsDtm[4170:5559, ]

# also save the labels
smsTrainy = smsRaw[1:4169, ]$type
smsTesty  = smsRaw[4170:5559, ]$type

> prop.table(table(smsTrainy))
smsTrainy
      ham      spam
0.8647158 0.1352842
> prop.table(table(smsTesty))
smsTesty
      ham      spam
0.8683453 0.1316547
```

Throw Away Terms with Low Frequency

```
# save frequently-appearing terms to a character vector
smsFreqWords = findFreqTerms(smsTrain, 5)

> str(smsFreqWords)
 chr [1:1136] "abiola" "abl" "abt" "accept" "access" "account" ...

> length(smsFreqWords)
[1] 1136

# create DTMs with only the frequent terms
smsFreqTrain = smsTrain[ , smsFreqWords]
smsFreqTest = smsTest[ , smsFreqWords]

> dim(smsFreqTrain)
[1] 4169 1136
> dim(smsTest)
[1] 1390 1136
```

Convert Counts to Indicators

Convert number of times a term is in a document to just whether or not it is in the document.

```
#convert counts to if(count>0) (yes,no)
convertCounts <- function(x) {
  x <- ifelse(x > 0, "Yes", "No")
}
# apply() convert_counts() to columns of train/test data
# these are just matrices
smsTrain = apply(smsFreqTrain, MARGIN = 2, convertCounts)
smsTest  <- apply(smsFreqTest, MARGIN = 2, convertCounts)

> dim(smsTrain)
[1] 4169 1136
> smsTrain[1:3,1:5]
    Terms
Docs abiola abl  abt  accept access
   1 "No"   "No" "No" "No"   "No"
   2 "No"   "No" "No" "No"   "No"
   3 "No"   "No" "No" "No"   "No"
```

We are ready for NB!!!

```
library(e1071)
smsNB = naiveBayes(smsTrain, smsTrainy)

> smsNB$tables[1:3]
$abiola
         abiola
smsTrainy          No         Yes
     ham  0.998058252 0.001941748
     spam 1.000000000 0.000000000

$abl
         abl
smsTrainy          No         Yes
     ham  0.994729542 0.005270458
     spam 1.000000000 0.000000000

$abt
         abt
smsTrainy          No         Yes
     ham  0.995839112 0.004160888
     spam 1.000000000 0.000000000
```

The tables are our $p(x_i|y)$ terms !!
y is ham or spam and x_i are the words(terms): *abiola*, *abl*, *abt*, ...

That is $p(abiola = Yes \,|\, y = ham) = 0.001941748$.

```
$age
         age
smsTrainy          No         Yes
     ham  0.998613037 0.001386963
     spam 0.978723404 0.021276596

$adult
         adult
smsTrainy          No         Yes
     ham  0.999445215 0.000554785
     spam 0.994680851 0.005319149
```

What is $p(y = spam|age = Yes)$?
(The prob the sms is spam given the word age is in it).

Let's use $p(y = spam) = .14$, the training data proportion.

$$p(y = spam|age = Yes) =$$

$$\frac{p(y{=}spam)p(age{=}Yes|y{=}spam)}{p(y{=}spam)p(age{=}Yes|y{=}spam)+p(y{=}ham)p(age{=}Yes|y{=}ham)}$$

```
> .14*0.021276596/(.14*0.021276596 + .86*0.001386963)
[1] 0.7140633
```

```
In [1]: priodds = .14/.86

In [2]: likerat = 0.021276596/0.001386963

In [3]: priodds
Out[3]: 0.16279069767441862

In [4]: likerat
Out[4]: 15.340420761044093

In [5]: postodds = priodds*likerat

In [6]: postodds
Out[6]: 2.497277798309504

In [7]: pspam = postodds/(1+postodds)

In [8]: pspam
Out[8]: 0.7140633207681201
```

age was 15 times more likely to be in the message if it was spam!!

```
$age
         age
smsTrainy          No         Yes
     ham  0.998613037 0.001386963
     spam 0.978723404 0.021276596

$adult
         adult
smsTrainy          No         Yes
     ham  0.999445215 0.000554785
     spam 0.994680851 0.005319149
```

What is $p(y = spam|age = Yes, adult = Yes)$?
(The prob the sms is spam given the word age is in it and the word adult is in it).

$$p(y = spam|age = Yes, adult = Yes) =$$

$$\frac{p(spam)p(age{=}Y|spam)p(adult{=}Y|spam)}{p(spam)p(age{=}Y|spam)p(adult{=}Y|spam)+p(ham)p(age{=}Y|ham)p(age{=}Y|ham)}$$

```
> .14*0.021276596*0.005319149/(.14*0.021276596*0.005319149 + .86*0.001386963*0.000554785)
[1] 0.9599091
```

Ok, let's try it with all the terms (words) !!!

Out of Sample Confusion Matrix

```
yhat = predict(smsNB,smsTest)

library(gmodels)
CrossTable(yhat, smsTesty,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
             | actual
   predicted |       ham |      spam | Row Total |
-------------|-----------|-----------|-----------|
         ham |      1201 |        30 |      1231 |
             |     0.995 |     0.164 |           |
-------------|-----------|-----------|-----------|
        spam |         6 |       153 |       159 |
             |     0.005 |     0.836 |           |
-------------|-----------|-----------|-----------|
Column Total |      1207 |       183 |      1390 |
             |     0.868 |     0.132 |           |
-------------|-----------|-----------|-----------|
```

Missclassification rate:

$36/1390 = 0.02589928$

% spam detected: .836.

```
> 153/(153+30)
[1] 0.8360656
```

Not bad !!

Try it again with laplace=1, add 1 to each count when estimating th 2x2 tables.

```
> smsNB2 = naiveBayes(smsTrain, smsTrainy,laplace=1)
> smsNB2$tables[1:3]
$abiola
         abiola
smsTrainy          No         Yes
     ham  0.997782090 0.002217910
     spam 0.998233216 0.001766784

$abl
         abl
smsTrainy          No         Yes
     ham  0.994455226 0.005544774
     spam 0.998233216 0.001766784

$abt
         abt
smsTrainy          No         Yes
     ham  0.995564181 0.004435819
     spam 0.998233216 0.001766784
```

Got rid of the 0/1 conditional probabilities, which are extreme.

```
> yhat2 = predict(smsNB2,smsTest)
> CrossTable(yhat2, smsTesty,
+            prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
+            dnn = c('predicted', 'actual'))


             | actual
   predicted |       ham |      spam | Row Total |
-------------|-----------|-----------|-----------|
         ham |      1202 |        28 |      1230 |
             |     0.996 |     0.153 |           |
-------------|-----------|-----------|-----------|
        spam |         5 |       155 |       160 |
             |     0.004 |     0.847 |           |
-------------|-----------|-----------|-----------|
Column Total |      1207 |       183 |      1390 |
             |     0.868 |     0.132 |           |
-------------|-----------|-----------|-----------|
```

Not too different.