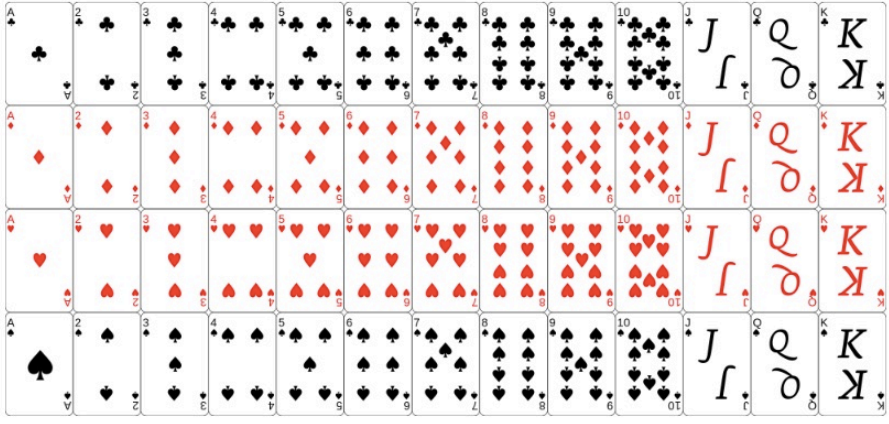
Naive Bayes

- Known for being simple but very efficient
- Makes one **very** strong assumption about the data
- High bias method
- For small training sets, outperforms much more sophisticated models
- Usually a good thing to try, just to see if it works

Probability Basics



- 52 cards total
- 2 colors (red and black)
- 4 suits (clubs, spaids, diamonds, hearts)
- A, 2, …, 10, J, K, Q. 13 possible values, one each per suit

Joint Probability and the Product Rule

Let V be a random variable that takes on card values

- e.g. $V = ACE$ or $V = 7$

Let C be a random variable that takes on *colors*

- e.g. $C = RED$ or $C = BLACK$

The joint probability of V and C, written $p(V, C)$, is the probability that card value V and color C appear simultaneously.

Example: The probability that we draw a RED 7 from the deck can be written as $p(V = 7, C = RED)$

Joint Probability and the Product Rule

Joint probabilities can be expressed via conditional probabilities

$$p(V, C) = p(V \mid C)p(C) = p(C \mid V)p(V)$$

This is sometimes called the **product rule**.

Example: Compute $p(V = 7, C = RED)$

We can get the joint probability fairly easily just from intuition

There are 2 red 7's in the deck (7 of Hearts and 7 of Diamonds),

$$p(V = 7, C = RED) = \frac{2}{52} = \frac{1}{26}$$

Joint Probability and the Product Rule

Joint probabilities can be expressed via conditional probabilities

$$p(V, C) = p(V \mid C)p(C) = p(C \mid V)p(V)$$

Example: Compute $p(V = 7, C = RED)$

But now we'll do it using the **product rule**:

$$p(V = 7 \mid C = RED)p(C = RED)$$

$$p(C = RED) = \frac{1}{2}, \;\; p(V = 7 \mid C = RED) = \frac{1}{13}$$

$$p(V = 7 \;\; C = RED) = \frac{1}{13} \cdot \frac{1}{2} = \frac{1}{26}$$

The Chain Rule of Probability

Say we have the joint prob. of 3 random variables: A, B, and C

By repeated use of the product rule, we can show

$$p(A, B, C) = p(A)\, p(B \mid A)\, p(C \mid A, B)$$

For D random variables $X_{1:D} = X_1, X_2, \ldots, X_D$ we have

$$p(X_{1:D}) = p(X_1)\, p(X_2 \mid X_1)\, p(X_3 \mid X_1, X_2) \cdots p(X_D \mid X_{1:D-1})$$

This is sometimes called the **chain rule** of probability

EFY: Use the **product rule** to prove the **chain rule** for the 3 random variable case above.

Marginal Probability and the Sum Rule

The marginal probability of V is just $p(V)$

If we know the joint probability of two random variables, we can compute the marginal for V as

$$p(V) = \sum_c p(V, C = c) = \sum_c p(V \mid C = c)p(C = c)$$

Sometimes called the **sum rule** or the **rule of total probability**

Example: Use the **sum rule** to evaluate the marginal $p(V = 7)$

Marginal Probability and the Sum Rule

$$p(V) = \sum_c p(V, C = c) = \sum_c p(V \mid C = c) p(C = c)$$

Example: Use the **sum rule** to evaluate the marginal $p(V = 7)$

There are two possible values for C (RED and BLACK), giving

$$\begin{aligned} p(V = 7) &= p(V = 7 \mid C = RED)\, p(C = RED) \\ &+ p(V = 7 \mid C = BLACK)\, p(C = BLACK) \\ &= \frac{1}{13} \cdot \frac{1}{2} + \frac{1}{13} \cdot \frac{1}{2} \\ &= \frac{1}{13} \end{aligned}$$

Bayes Rule

Note that we can rewrite the **product rule** and get

$$p(Y \mid X) = \frac{p(Y, X)}{p(X)}, \quad \text{if } p(X) > 0$$

And using the **product rule** again on the numerator, we have

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)}$$

which is the classical statement of **Bayes Rule**

... but let's go a little further

Bayes Rule

Say we evaluate the conditional prob. for values $X = x$ and $Y = y$

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

and then use the **sum rule** on the denominator

$$p(Y = y \mid X = x) = \frac{p(X = x \mid Y = y)\, p(Y = y)}{\sum_{y'} p(X = x \mid Y = y')\, p(Y = y')}$$

Note that this allows us to compute $P(Y \mid X)$ using only conditionals of the form $P(X \mid Y)$ and the marginal for Y

Bayes Rule - Classic Cancer Test Example

Let's assume we know that 1% of women over the age of 40 have breast cancer

$$p(\mathrm{C}) = 0.01$$

Let's assume that 90% of women who **have cancer** will test positive for cancer in a mammogram.

$$p(\mathrm{pos} \mid \mathrm{C}) = 0.90$$

Finally, assume that 8% of women that do **not** have cancer will also test positive

$$p(\mathrm{pos} \mid \mathrm{not\ C}) = 0.08$$

Bayes Rule - Classic Cancer Test Example

What is the probability that a woman who tests positive for cancer **actually has cancer**? In other words, what is

$$p(\mathrm{C} \mid \mathrm{pos})$$

Most people will assume that if they get a positive test, then there is a 90% chance that they actually have cancer.

But this ignores the incredibly important fact that **not many people have cancer**! Remember:

$$p(\mathrm{C}) = 0.01$$

Bayes Rule - Classic Cancer Test Example

Let's do the actual calculation. From Bayes Law, we have

$$p(\text{C} \mid \text{pos}) = \frac{p(\text{pos} \mid \text{C})\, p(\text{C})}{p(\text{pos} \mid \text{C})\, p(\text{C}) + p(\text{pos} \mid \text{not C})\, p(\text{not C})}$$

The only quantity we haven't specified is the probability of not having cancer

$$p(\text{not C}) = 1 - p(\text{C}) = 1 - 0.01 = 0.99$$

Plugging everything into Bayes Law gives ...

Bayes Rule - Classic Cancer Test Example

Let's do the actual calculation. From Bayes Law, we have

$$p(\mathrm{C} \mid \mathrm{pos}) = \frac{0.90 \cdot 0.01}{0.90 \cdot 0.01 + 0.08 \cdot 0.99} = 0.10$$

So even if you test positive for cancer, there is only about a 10% chance that you actually **have** cancer.

Naive Bayes

We said that we were going to model the joint probability $p(\mathbf{x}, y)$

But really we'll use the product rule first: $p(\mathbf{x}, y) = p(\mathbf{x} \mid y)\, p(y)$

We still want to get at $p(y \mid \mathbf{x})$

Which we can do with Bayes Law $p(y \mid \mathbf{x}) = \dfrac{p(\mathbf{x} \mid y)\, p(y)}{p(\mathbf{x})}$

Stated another way: $\text{posterior} = \dfrac{\text{class-conditional} \cdot \text{prior}}{\text{evidence}}$

Naive Bayes

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

$$\text{posterior} = \frac{\text{class-conditional} \cdot \text{prior}}{\text{evidence}}$$

Posterior Probability $p(y \mid \mathbf{x})$

Can be interpretted as asking:

"What is the probability that a particular object belongs to class c given its observed features"

Or in concrete terms:

"What is the probability that an email is spam given its content?"

Given an email $\mathbf{x}$ we want to classify:

email is spam if
$p(\text{spam} \mid \mathbf{x}) \geq p(\text{ham} \mid \mathbf{x})$
else classify email as ham

Naive Bayes

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

$$\text{posterior} = \frac{\text{class-conditional} \cdot \text{prior}}{\text{evidence}}$$

Class-Conditional Probability $p(\mathbf{x} \mid y)$

Sometimes called the **likelihood**:

"Given a class $y = c$, what is the probability that $\mathbf{x}$ is observed?"

Or more concretely:

"Given assumptions made about the nature of spam email, what are the chances I would get *this* email?"

Example: $p(\mathbf{x} = [\text{buy}, \text{viagra}] \mid y = spam)$

Joint probability of features is hard to estimate

Here is where we make our **naive** assumption

Class-Conditional Probability $p(\mathbf{x} \mid y)$

Assumption: Features of $\mathbf{x}$ are conditionally independent given the class y

Example: Two (possibly) differently waited coins, C_1 and C_2. Pick a coin and flip it three times.

$$p(\mathbf{x} = [\mathrm{H\ H\ T}] \mid C_1) = p(\mathrm{H} \mid C_1) \cdot p(\mathrm{H} \mid C_1) \cdot p(\mathrm{T} \mid C_1)$$

In this case, conditional independence of the three coin flips is actually valid.

Class-Conditional Probability $p(\mathbf{x} \mid y)$

Assumption: Features of $\mathbf{x}$ are conditionally independent given the class y

Example: A particular spam email

$$p(\mathbf{x} = [\text{buy, viagra, deal}] \mid \text{spam}) =$$

$$p(\text{buy} \mid \text{spam}) \cdot p(\text{viagra} \mid \text{spam}) \cdot p(\text{deal} \mid \text{spam})$$

Is this a valid assumption?

Probably not, but we're going to make it anyway because it makes the feature conditionals super easy to estimate from the data

Class-Conditional Probability $p(\mathbf{x} \mid y)$

Assumption: Features of $\mathbf{x}$ are conditionally independent given the class y

Example: A particular spam email

$$p(\mathbf{x} = [\text{buy, viagra, deal}] \mid \text{spam}) =$$

$$p(\text{buy} \mid \text{spam}) \cdot p(\text{viagra} \mid \text{spam}) \cdot p(\text{deal} \mid \text{spam})$$

Is this a valid assumption?

Example: $\hat{p}(\text{deal} \mid \text{spam}) = \dfrac{\#\text{ deal in spam messages}}{\#\text{ words in spam messages}}$

Naive Bayes

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

$$\text{posterior} = \frac{\text{class-conditional} \cdot \text{prior}}{\text{evidence}}$$

Prior Probability $p(y)$

Sometimes called the **class prior probability**

"the general probability of encountering a particular class"

Or concretely:

$p(\text{spam})$ = "the probability that any new message is a spam"

How do we get the prior?

Ask a subject-matter expert

- Experts believe that 80% of all email is spam

Prior Probability $p(y)$

Sometimes called the **class prior probability**

"the general probability of encountering a particular class"

Or concretely:

$p(\text{spam})$ = "the probability that any new message is a spam"

How do we get the prior?

Estimate it from the Data

$$\hat{p}(\text{spam}) = \frac{\#\text{ of messages that are spam}}{\#\text{ of messages}}$$

Naive Bayes

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

$$\text{posterior} = \frac{\text{class-conditional} \cdot \text{prior}}{\text{evidence}}$$

Evidence $p(\mathbf{x})$

"The probability of encountering data $\mathbf{x}$ independent of class label"

Concretely:

"The probability of receiving message $\mathbf{x}$ whether it's spam or ham"

Could compute using the **sum rule**

But we won't because it doesn't actually help us make descisions

$$\frac{p(\mathbf{x} \mid \text{spam}) \cdot p(\text{spam})}{p(\mathbf{x})} \geq \frac{p(\mathbf{x} \mid \text{ham}) \cdot p(\text{ham})}{p(\mathbf{x})}$$

Denominator is same in both

Evidence $p(\mathbf{x})$

"The probability of encountering data $\mathbf{x}$ independent of class label"

Concretely:

"The probability of receiving message $\mathbf{x}$ whether it's spam or ham"

Could compute using the **sum rule**

But we won't because it doesn't actually help us make descisions

$$p(\mathbf{x} \mid \text{spam}) \cdot p(\text{spam}) \geq p(\mathbf{x} \mid \text{ham}) \cdot p(\text{ham})$$

Can't think of as probabilities anymore. Better to think of as **scores**.

Spam vs. Ham Example

Example: Compute the ham score for $\mathbf{x}$ = [fly , nigeria]

ham	*spam*	*spam*	*spam*	*ham*
work	*nigeria*	*fly*	*money*	*fly*
buy	*opportunity*	*buy*	*buy*	*home*
money	*viagra*	*nigeria*	*fly*	*nigeria*

$$p(\mathbf{x} \mid \text{ham})p(\text{ham}) =$$

$$p(\text{ fly } \mid \text{ham})\, p(\text{nigeria} \mid \text{ham})\, p(\text{ham}) =$$

$$\frac{1}{6} \qquad \frac{1}{6} \qquad \frac{2}{5} \quad = \quad \frac{1}{90}$$

Spam vs. Ham Example

Example: Compute the spam score for $\mathbf{x}$ = [fly , nigeria]

ham	*spam*	*spam*	*spam*	*ham*
work	*nigeria*	*fly*	*money*	*fly*
buy	*opportunity*	*buy*	*buy*	*home*
money	*viagra*	*nigeria*	*fly*	*nigeria*

$$p(\mathbf{x} \mid \text{spam})p(\text{spam}) =$$

$$p(\text{ fly } \mid \text{spam})\, p(\text{nigeria} \mid \text{spam})\, p(\text{spam}) =$$

$$\frac{2}{9} \qquad \frac{2}{9} \qquad \frac{3}{5} \quad = \quad ??$$