University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chris Ketelsen

Lecture 2: Naive Bayes

Probabilistic Classification

Given: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ training examples

$$\mathbf{x}_i \in \mathbb{R}^D \qquad y_i \in \{c_1, c_2, \ldots, c_J\}$$

Goal: Given new data $\mathbf{x}$, predict its label y

For each class c_j, estimate

$$p(y = c_j \mid \mathbf{x}, \mathcal{D})$$

Assign to $\mathbf{x}$ the class with highest probability

$$\hat{y} = \arg\max_c p(y = c \mid \mathbf{x}, \mathcal{D})$$

Generative vs. Discriminative Models

How do we model/estimate these conditional probabilities?

Generative:

- Model the joint probability distribution $p(\mathbf{x}, y)$.
- Make assumptions about relationship between $\mathbf{x}$ and y
- Make assumptions about data ($\mathbf{x}$) itself
- **Today**: Naive Bayes

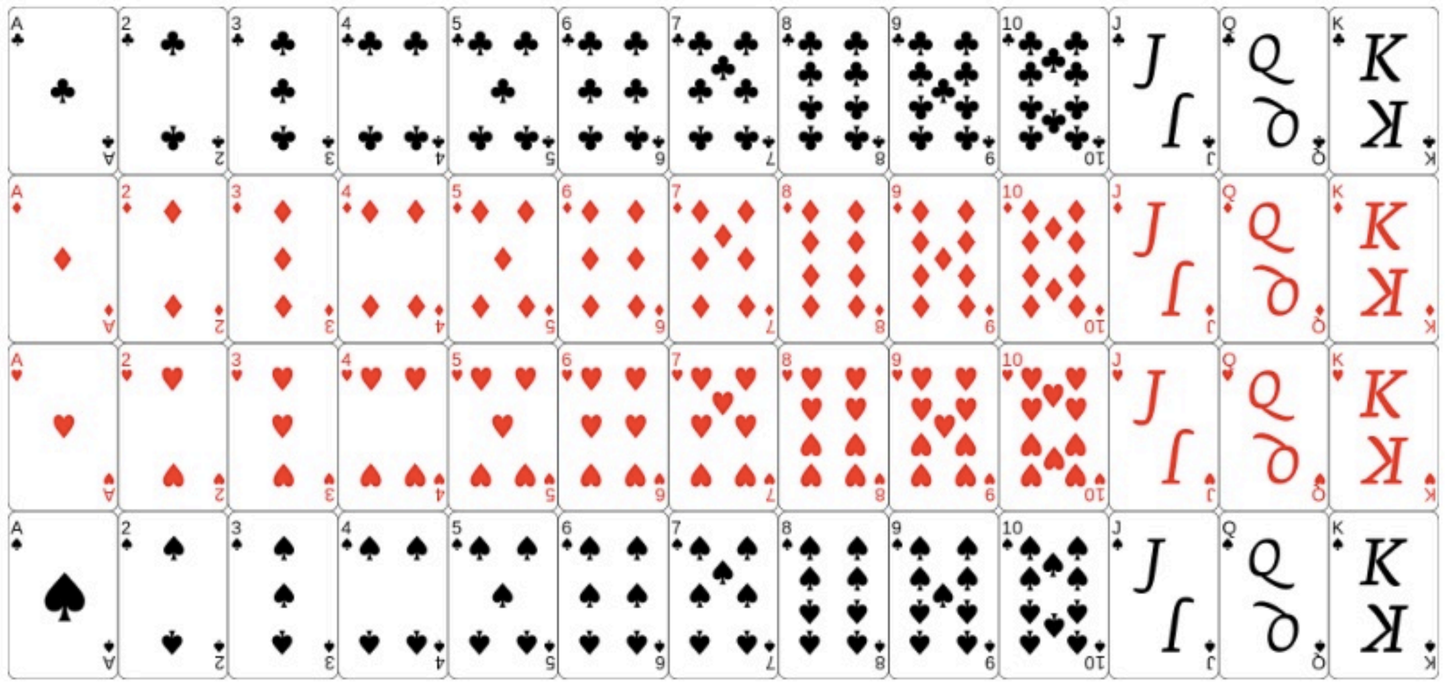
Discriminative:

- Model only conditional relationship $p(y \mid \mathbf{x})$
- **Next Time**: Logistic Regression

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**:

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

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

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

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

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

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

.90 .01

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

.90 .01 .09 .99

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

EMAIL

SPAM/HAM

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.

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

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

MAXIMUM LIKELIHOOD ESTIMATE

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

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

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

$p([\text{buy}, \text{money}, \text{viagra}])$

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 $\mathbf{x}$ = [work, nigeria]

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

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

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

$\frac{1}{6} \cdot \frac{1}{6} \cdot \frac{2}{5} = \frac{1}{90}$

Spam vs. Ham Example

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

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

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

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

$\frac{0}{9} \cdot \frac{2}{9} \cdot \frac{3}{5} = 0$

Spam vs. Ham Example

We found that $p(\text{spam} \mid \mathbf{x} = [\text{work, nigeria}]) = 0$

This should bother you...

The problem is that "work" did not show up in any training examples labeled spam.

From the spammer's perspective, including an uncommon word not in the spam training set, would lead to $p(\text{spam} \mid \mathbf{x}) = 0$ and the filter would never catch it.

It would be nice if we could ensure that no messages result in a zero probability.

Additive Smoothing

Want to avoid zero probabilities when word w isn't in training set

Silly idea: Add 1 to all of the word counts

$$\hat{p}(\mathrm{w} \mid \mathrm{class}) = \frac{\#\ \mathrm{w\ in\ class} + 1}{\#\ \mathrm{words\ in\ class}}$$???

Nice if $\hat{p}(\mathrm{w} \mid \mathrm{class})$ behaved like a probability distribution

w can take on any value in the vocabulary (call it V)

restore sum-to-one by adding number of words in the vocabular to denominator

Additive Smoothing

Want to avoid zero probabilities when word w isn't in training set

$$\hat{p}(\text{w} \mid \text{class}) = \frac{\#\text{ w in class} + 1}{\#\text{ words in class} + |\text{V}|}$$

What about the (ridiculous) case when we have no spam documents or no ham documents in the training set?

$$\hat{p}(\text{class}) = \frac{\#\text{ of messages that are in class} + 1}{\#\text{ of messages} + |\text{C}|}$$

where $|C|$ is the number of classes (always 2 for spam vs ham)

add-1 smoothing is called **Laplace Smoothing**

Spam vs. Ham Example Revisited

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

$|V| = 8$

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

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

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

$$\frac{0+1}{9+8} \cdot \frac{2+1}{9+8} \cdot \frac{3+1}{5+2}$$

$$= \frac{1}{17} \cdot \frac{3}{17} \cdot \frac{4}{7} = 0.00346$$

Spam vs. Ham Example Revisited

.003 vs .009
SPAM HAM

EFY: Compute the ham score for $\mathbf{x}$ = [work, nigeria] with smoothing

|V| = 8

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

⇒ HAM

$P(\text{WORK} \mid \text{HAM}) \cdot P(\text{NIGERIA} \mid \text{HAM}) \, P(\text{HAM})$

$P(x|c)\,p(c)$ / ~~$p(x)$~~

$\frac{1+1}{6+8} \cdot \frac{1+1}{6+8} \cdot \frac{2+1}{5+2}$

$\frac{2}{14} \cdot \frac{2}{14} \cdot \frac{3}{7} = .000875$

Naive Bayes Classifier

- The Naive Bayes Classifier is a probabilistic classifier
- We compute the probability of message $\mathbf{x}$ of length D being in class c as

$$p(c \mid \mathbf{x}) \propto p(c) \prod_{i=1}^{D} p(x_i \mid c)$$

- Predicted class is the one with the highest posterior score

$$\hat{y} = \arg\max_c \hat{p}(c \mid \mathbf{x}) = \arg\max_c \hat{p}(c) \prod_{i=1}^{D} \hat{p}(x_i \mid c)$$

Numerical Hiccup

For large messages, have to multiply a lot of probabilities together

$$\hat{y} = \arg\max_c \hat{p}(c) \prod_{i=1}^{D} \hat{p}(x_i \mid c)$$

Probabilities are ≤ 1. If you multiply enough of them together you could run into **underflow**

Fix: Maximize the log instead

Recall: $\ln(ab) = \ln(a) + \ln(b)$

Important: Since $\ln(x)$ is a monotonically increasing function, the maximizer does not change

Numerical Hiccup

Problem becomes

$$\hat{y} = \arg\max_c \ \ln \hat{p}(c) + \sum_{i=1}^{D} \ln \hat{p}(x_i \mid c)$$