Classification I: Naïve Bayes and Logistic Regression

Digging into Data: Jordan Boyd-Graber

University of Maryland

February 18, 2013

Slides adapted from Hinrich Schütze and Lauren Hannah

Roadmap

- Classification
- Estimating probability distributions
- Naïve Bayes Example
- Logistic regression
- Evaluating classification

Outline

Formal definition of Classification

Given:

- A universe $\mathbb{X}$ our examples can come from (e.g., English documents with a predefined vocabulary)

Formal definition of Classification

Given:

- A universe $\mathbb{X}$ our examples can come from (e.g., English documents with a predefined vocabulary)
  - Examples are represented in this space. (e.g., each document has some subset of the vocabulary; more in a second)

Formal definition of Classification

Given:

- A universe $\mathbb{X}$ our examples can come from (e.g., English documents with a predefined vocabulary)
  - Examples are represented in this space. (e.g., each document has some subset of the vocabulary; more in a second)
- A fixed set of classes $\mathbb{C} = \{c_1, c_2, \dots, c_J\}$

Formal definition of Classification

Given:

- A universe $\mathbb{X}$ our examples can come from (e.g., English documents with a predefined vocabulary)
  - Examples are represented in this space. (e.g., each document has some subset of the vocabulary; more in a second)
- A fixed set of classes $\mathbb{C} = \{c_1, c_2, \dots, c_J\}$
  - The classes are human-defined for the needs of an application (e.g., spam vs. ham).

Formal definition of Classification

Given:

- A universe $\mathbb{X}$ our examples can come from (e.g., English documents with a predefined vocabulary)
  - Examples are represented in this space. (e.g., each document has some subset of the vocabulary; more in a second)
- A fixed set of classes $\mathbb{C} = \{c_1, c_2, \dots, c_J\}$
  - The classes are human-defined for the needs of an application (e.g., spam vs. ham).
- A training set D of labeled documents with each labeled document $d \in \mathbb{X} \times \mathbb{C}$

Formal definition of Classification

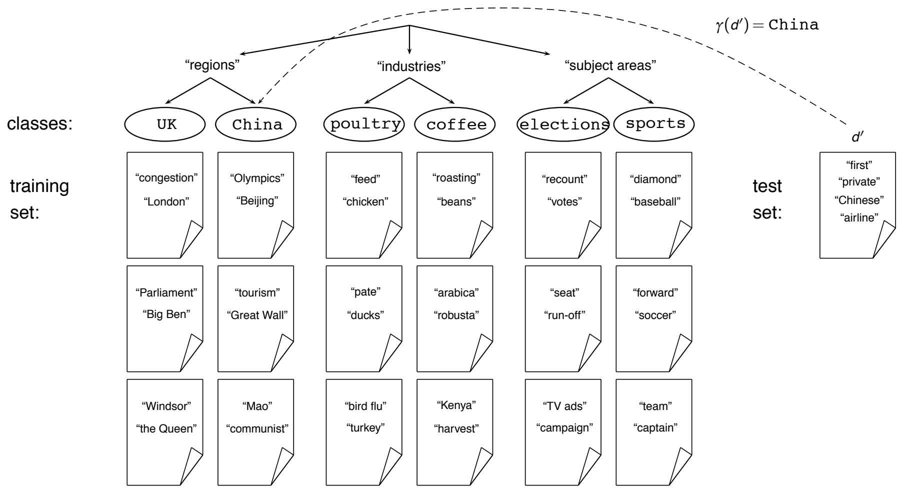
Given:

- A universe $\mathbb{X}$ our examples can come from (e.g., English documents with a predefined vocabulary)
  - Examples are represented in this space. (e.g., each document has some subset of the vocabulary; more in a second)
- A fixed set of classes $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$
  - The classes are human-defined for the needs of an application (e.g., spam vs. ham).
- A training set D of labeled documents with each labeled document $d \in \mathbb{X} \times \mathbb{C}$

Using a learning method or learning algorithm, we then wish to learn a classifier γ that maps documents to classes:

$$\gamma : \mathbb{X} \rightarrow \mathbb{C}$$

Topic classification



Examples of how search engines use classification

- Standing queries (e.g., Google Alerts)
- Language identification (classes: English vs. French etc.)
- The automatic detection of spam pages (spam vs. nonspam)
- The automatic detection of sexually explicit content (sexually explicit vs. not)
- Sentiment detection: is a movie or product review positive or negative (positive vs. negative)
- Topic-specific or *vertical* search – restrict search to a “vertical” like “related to health” (relevant to vertical vs. not)

Classification methods: 1. Manual

- Manual classification was used by Yahoo in the beginning of the web. Also: ODP, PubMed
- Very accurate if job is done by experts
- Consistent when the problem size and team is small
- Scaling manual classification is difficult and expensive.
- $\longrightarrow$ We need automatic methods for classification.

Classification methods: 2. Rule-based

- There are “IDE” type development enviroments for writing very complex rules efficiently. (e.g., Verity)
- Often: Boolean combinations (as in Google Alerts)
- Accuracy is very high if a rule has been carefully refined over time by a subject expert.
- Building and maintaining rule-based classification systems is expensive.

Classification methods: 3. Statistical/Probabilistic

- As per our definition of the classification problem – text classification as a learning problem
- Supervised learning of a the classification function γ and its application to classifying new documents
- We will look at a couple of methods for doing this: Naive Bayes, Rocchio, kNN
- No free lunch: requires hand-classified training data
- But this manual classification can be done by non-experts.

Outline

The Naive Bayes classifier

- The Naive Bayes classifier is a probabilistic classifier.
- We compute the probability of a document d being in a class c as follows:

$$P(c|d) \propto P(c) \prod_{1 \leq i \leq n_d} P(w_i|c)$$

- n_d is the length of the document. (number of tokens)
- $P(w_i|c)$ is the conditional probability of term w_i occurring in a document of class c
- $P(w_i|c)$ as a measure of how much evidence w_i contributes that c is the correct class.
- $P(c)$ is the prior probability of c.
- If a document's terms do not provide clear evidence for one class vs. another, we choose the c with higher $P(c)$.

Maximum a posteriori class

- Our goal is to find the “best” class.
- The best class in Naive Bayes classification is the most likely or **m**aximum a posteriori (MAP) class c_{map} :

$$c_{\text{map}} = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j|d) = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j) \prod_{1 \leq i \leq n_d} \hat{P}(w_i|c_j)$$

- We write $\hat{P}$ for P since these values are **e**stimates from the training set.

Outline

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{``}buy\text{''} | z = \mathrm{SPAM})$.

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{``}buy\text{''} | z = \texttt{SPAM})$.

buy	**buy**	nigeria	opportunity	viagra
nigeria	opportunity	viagra	fly	money
fly	**buy**	nigeria	fly	**buy**
money	**buy**	fly	nigeria	viagra

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{``}buy\text{''} | z = \texttt{SPAM})$.

buy	**buy**	nigeria	opportunity	viagra
nigeria	opportunity	viagra	fly	money
fly	**buy**	nigeria	fly	**buy**
money	**buy**	fly	nigeria	viagra

- Maximum likelihood (ML) estimate of the probability is:

$$\hat{\theta}_i = \frac{n_i}{\sum_k n_k} \tag{1}$$

How do we estimate a probability?

- Suppose we want to estimate $P(w_n = \text{``}buy\text{''} | z = \texttt{SPAM})$.

buy	**buy**	nigeria	opportunity	viagra
nigeria	opportunity	viagra	fly	money
fly	**buy**	nigeria	fly	**buy**
money	**buy**	fly	nigeria	viagra

- Maximum likelihood (ML) estimate of the probability is:

$$\hat{\theta}_i = \frac{n_i}{\sum_k n_k} \tag{1}$$

- Is this reasonable?

The problem with maximum likelihood estimates: Zeros (cont)

- If there were no occurrences of "bagel" in documents in class SPAM, we'd get a zero estimate:

$$\hat{P}(\text{“bagel”} | \texttt{SPAM}) = \frac{T_{\texttt{SPAM}, \text{“bagel”}}}{\sum_{w' \in V} T_{\texttt{SPAM}, w'}} = 0$$

- $\rightarrow$ We will get $P(\texttt{SPAM}|d) = 0$ for any document that contains bagel!
- Zero probabilities cannot be conditioned away.

How do we estimate a probability?

- In computational linguistics, we often have a *prior* notion of what our probability distributions are going to look like (for example, non-zero, sparse, uniform, etc.).
- This estimate of a probability distribution is called the maximum a posteriori (MAP) estimate:

$$\theta_{\text{MAP}} = \text{argmax}_{\theta} f(x|\theta)g(\theta) \tag{2}$$

How do we estimate a probability?

- For a multinomial distribution (i.e. a discrete distribution, like over words):

$$\theta_i = \frac{n_i + \alpha_i}{\sum_k n_k + \alpha_k} \tag{3}$$

- α_i is called a smoothing factor, a pseudocount, etc.

How do we estimate a probability?

- For a multinomial distribution (i.e. a discrete distribution, like over words):

$$\theta_i = \frac{n_i + \alpha_i}{\sum_k n_k + \alpha_k} \tag{3}$$

- α_i is called a smoothing factor, a pseudocount, etc.
- When $\alpha_i = 1$ for all i, it's called “Laplace smoothing” and corresponds to a uniform prior over all multinomial distributions (just do this).

How do we estimate a probability?

- For a multinomial distribution (i.e. a discrete distribution, like over words):

$$\theta_i = \frac{n_i + \alpha_i}{\sum_k n_k + \alpha_k} \tag{3}$$

- α_i is called a smoothing factor, a pseudocount, etc.
- When $\alpha_i = 1$ for all i, it's called "Laplace smoothing" and corresponds to a uniform prior over all multinomial distributions (just do this).
- To geek out, the set $\{\alpha_1, \dots, \alpha_N\}$ parameterizes a Dirichlet distribution, which is itself a distribution over distributions and is the conjugate prior of the Multinomial (don't need to know this).

Outline

A Classification Problem

- Suppose that I have two coins, C_1 and C_2
- Now suppose I pull a coin out of my pocket, flip it a bunch of times, record the coin and outcomes, and repeat many times:

```
C1: 0 1 1 1 1
C1: 0 1 0
C2: 1 0 0 0 0 0 0 1
C1: 0 1
C1: 1 1 0 1 1 1
C2: 0 0 1 1 0 1
C2: 1 0 0 0
```

- Now suppose I am given a new sequence, `0 0 1 0 0 1`; which coin is it from?

A Classification Problem

This problem has particular challenges:

- different numbers of covariates for each observation
- number of covariates can be large

K-nn will not work: what is the distance between `0 1 0` and `1 0 0 0 0 0 0 1`?

However, there is some structure:

- Easy to get $P(C_1)$, $P(C_2)$
- Also easy to get $P(X_i = 1 | C_1)$ and $P(X_i = 1 | C_2)$
- By conditional independence,

$$P(X = 0\,1\,0 \,|\, C_1) = P(X_1 = 0 | C_1) P(X_2 = 1 | C_1) P(X_2 = 0 | C_1)$$

- Can we use these to get $P(C_1 | X = 0\,0\,1\,0\,0\,1)$?

A Classification Problem

Summary: have $P(data|class)$, want $P(class|data)$

Solution: Bayes' rule!

$$P(class|data) = \frac{P(data|class)P(class)}{P(data)} = \frac{P(data|class)P(class)}{\sum_{class=1}^{C} P(data|class)P(class)}$$

To compute, we need to estimate $P(data|class)$, $P(class)$ for all classes

A Classification Problem

Training data:

```
C1: 0 1 1 1 1
C1: 0 1 0
C2: 1 0 0 0 0 0 0 1
C1: 0 1
C1: 1 1 0 1 1 1
C2: 0 0 1 1 0 1
C2: 1 0 0 0
```

Testing data: `0 0 1 0 0 1`

Estimate $P(C_1)$, $P(C_2)$, $P(X = 001001 \,|\, C_1)$, $P(X = 001001 \,|\, C_2)$

Naive Bayes Classifier

This works because the coin flips are independent given the coin parameter. What about this case:

- want to identify the type of fruit given a set of features: color, shape and size
- color: red, green, yellow or orange (categorical)
- shape: round, oval or long+skinny (categorical)
- size: diameter in inches (continuous)

Naive Bayes Classifier

Conditioned on type of fruit, these features are not necessarily independent:

Given category "apple," the color "green" has a higher probability given "size < 2":

$$P(green \mid size < 2, apple) > P(green \mid apple)$$

Naive Bayes Classifier

Using chain rule,

$$\begin{aligned}
&P(apple \mid green, round, size = 2) \\
&\qquad = \frac{P(green, round, size = 2 \mid apple) P(apple)}{\sum_{fruits} P(green, round, size = 2 \mid fruit\, j) P(fruit\, j)} \\
&\qquad \propto P(green \mid round, size = 2, apple) P(round \mid size = 2, apple) \\
&\qquad\qquad \times P(size = 2 \mid apple) P(apple)
\end{aligned}$$

But computing conditional probabilities is hard! There are many combinations of $(color, shape, size)$ for each fruit.

Outline

Naive Bayes Classifier

Idea: assume conditional independence for all features given class,

$$P(green \mid round, size = 2, apple) = P(green \mid apple)$$
$$P(round \mid green, size = 2, apple) = P(round \mid apple)$$
$$P(size = 2 \mid green, round, apple) = P(size = 2 \mid apple)$$

More generally for features (or covariates) $X_1, \ldots, X_m$ and class Y,

$$P(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m, Y) = P(X_i \mid Y),$$
$$P(X_1, \ldots, X_m \mid Y) = \prod_{i=1}^{m} P(X_i \mid Y)$$

Naive Bayes Classifier

Why conditional independence?

- estimating multivariate functions (like $P(X_1, \ldots, X_m \,|\, Y)$) is mathematically hard, while estimating univariate ones is easier (like $P(X_i \,|\, Y)$)
- need less data to fit univariate functions well
- univariate estimators differ much less than multivariate estimator (low variance)
- ... but they may end up finding the wrong values (more bias)

Naïve Bayes conditional independence assumption

To reduce the number of parameters to a manageable size, we make the **N**aive Bayes conditional independence assumption :

$$P(d|c_j) = P(\langle w_1, \ldots, w_{n_d} \rangle | c_j) = \prod_{1 \le i \le n_d} P(X_i = w_i | c_j)$$

We assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(X_i = w_i | c_j)$.

Recall from earlier the estimates for these priors and conditional probabilities: $\hat{P}(c_j) = \frac{N_c}{N}$ and $\hat{P}(w|c) = \frac{T_{cw}+1}{(\sum_{w' \in V} T_{cw'})+B}$

Outline

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Data Set

ID	Color	Origin	Organic	Apple?
1	Red	Domestic	Conventional	Yes
2	Red	Domestic	Conventional	No
3	Red	Domestic	Conventional	Yes
4	Yellow	Domestic	Conventional	No
5	Yellow	Domestic	Organic	Yes
6	Yellow	Imported	Organic	No
7	Yellow	Imported	Organic	Yes
8	Yellow	Imported	Conventional	No
9	Red	Imported	Organic	No
10	Red	Domestic	Organic	Yes

Let's say we wanted to decide whether a conventional imported red fruit is likely to be an apple or not. We need to compute some probabilities! Which ones?

Computing Probabilities

$$p(\text{Apple}) = \frac{5+1}{10+2} = .5$$

$$p(\neg\text{Apple}) = \frac{5+1}{10+2} = .5$$

$$p(\text{Red}|\text{Apple})\frac{3+1}{5+2} = 0.57$$
$$p(\text{Imp}|\text{Apple})\frac{1+1}{5+2} = 0.29$$
$$p(\text{Conv}|\text{Apple})\frac{2+1}{5+2} = 0.43$$

$$p(\text{Red}|\neg\text{Apple})\frac{2+1}{5+2} = 0.43$$
$$p(\text{Imp}|\neg\text{Apple})\frac{3+1}{5+2} = 0.57$$
$$p(\text{Conv}|\neg\text{Apple})\frac{3+1}{5+2} = 0.57$$

Computing Probabilities

$p(\text{Apple}) = \frac{5+1}{10+2} = .5$

$p(\neg\text{Apple}) = \frac{5+1}{10+2} = .5$

$$p(\text{Red}|\text{Apple})\frac{3+1}{5+2} = 0.57$$
$$p(\text{Imp}|\text{Apple})\frac{1+1}{5+2} = 0.29$$
$$p(\text{Conv}|\text{Apple})\frac{2+1}{5+2} = 0.43$$

$$p(\text{Red}|\neg\text{Apple})\frac{2+1}{5+2} = 0.43$$
$$p(\text{Imp}|\neg\text{Apple})\frac{3+1}{5+2} = 0.57$$
$$p(\text{Conv}|\neg\text{Apple})\frac{3+1}{5+2} = 0.57$$

$$\begin{aligned} & p(\text{C}|\text{App})p(\text{I}|\text{App})p(\text{R}|\text{App})p(\text{Apple}) \\ = & .43 \cdot .29 \cdot .57 \cdot .5 \\ = & .035 \end{aligned}$$

$$\begin{aligned} & p(\text{C}|\neg\text{App})p(\text{I}|\neg\text{App})p(\text{R}|\neg\text{App})p(\neg\text{App}) \\ = & .57 \cdot .57 \cdot .43 \cdot .5 \\ = & .070 \end{aligned}$$

Computing Probabilities

$p(\text{Apple}) = \frac{5+1}{10+2} = .5$

$p(\neg\text{Apple}) = \frac{5+1}{10+2} = .5$

$$p(\text{Red}|\text{Apple})\frac{3+1}{5+2} = 0.57$$
$$p(\text{Imp}|\text{Apple})\frac{1+1}{5+2} = 0.29$$
$$p(\text{Conv}|\text{Apple})\frac{2+1}{5+2} = 0.43$$

$$p(\text{Red}|\neg\text{Apple})\frac{2+1}{5+2} = 0.43$$
$$p(\text{Imp}|\neg\text{Apple})\frac{3+1}{5+2} = 0.57$$
$$p(\text{Conv}|\neg\text{Apple})\frac{3+1}{5+2} = 0.57$$

$$\begin{aligned} & p(\text{C}|\text{App})p(\text{I}|\text{App})p(\text{R}|\text{App})p(\text{Apple}) \\ = & .43 \cdot .29 \cdot .57 \cdot .5 \\ = & .035 \end{aligned}$$

So, probably not an apple!

$$\begin{aligned} & p(\text{C}|\neg\text{App})p(\text{I}|\neg\text{App})p(\text{R}|\neg\text{App})p(\neg\text{App}) \\ = & .57 \cdot .57 \cdot .43 \cdot .5 \\ = & .070 \end{aligned}$$

Taking the log

- Multiplying lots of small probabilities can result in floating point underflow.
- Since $\log(xy) = \log(x) + \log(y)$, we can sum log probabilities instead of multiplying probabilities.
- Since log is a monotonic function, the class with the highest score does not change.
- So what we usually compute in practice is:

$$c_{\text{map}} = \arg\max_{c_j \in \mathbb{C}} [\hat{P}(c_j) \prod_{1 \le i \le n_d} \hat{P}(w_i|c_j)]$$

$$\arg\max_{c_j \in \mathbb{C}} [\log \hat{P}(c_j) + \sum_{1 \le i \le n_d} \log \hat{P}(w_i|c_j)]$$

Outline

Generative vs. Discriminative Models

- Easy to compute $p(x|y)$, we want $p(y|x)$
- Naïve Bayes used Bayes rule to reverse conditioning
- **Generative** models try to model both y and x
- **Discriminative** models only model y (what you care about) given x
- We'll talk about one model that does that called logistic regression
  - Logistic: A special mathematical function it uses
  - Regression: Combines a weight vector with observations to create an answer
- Naïve Bayes is a special case of logistic regression

Logistic Regression: Definition

- Weight vector w_i
- Observations X_i
- "Bias" w_0 (like background probabilities in naïve Bayes)

$$P(Y=0|X) = \frac{1}{1+\exp\left[w_0+\sum_i w_i X_i\right]} \tag{4}$$

$$P(Y=1|X) = \frac{\exp\left[w_0+\sum_i w_i X_i\right]}{1+\exp\left[w_0+\sum_i w_i X_i\right]} \tag{5}$$

- Math is much hairier! (See optional reading)
- For shorthand, we'll say that

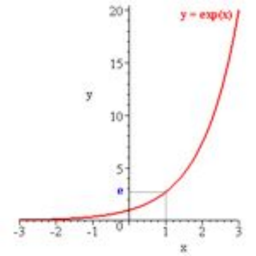
$$P(Y=0|X) = \sigma(-(w_0+\sum_i w_i X_i)) \tag{6}$$

$$P(Y=1|X) = 1-\sigma(-(w_0+\sum_i w_i X_i)) \tag{7}$$

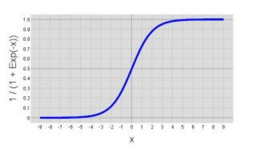
- Where $\sigma(z) = \frac{1}{1-exp[-z]}$

What's this "exp"?

Exponential



Logistic



- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - e^x is the limit of compound interest formula as compounds become infinitely small
  - It's the function whose derivative is itself
- The "logistic" function is $\sigma(z) = \frac{1}{1+e^{-z}}$
- Looks like an "S"
- Always between 0 and 1. Why is this useful?

Outline

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
“viagra”	w_1	2.0
“mother”	w_2	-1.0
“work”	w_3	-0.5
“nigeria”	w_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
"viagra"	w_1	2.0
"mother"	w_2	-1.0
"work"	w_3	-0.5
"nigeria"	w_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1]} = 0.48$
- $P(Y = 1) = \frac{\exp[0.1]}{1+\exp[0.1]} = .52$

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
“viagra”	w_1	2.0
“mother”	w_2	-1.0
“work”	w_3	-0.5
“nigeria”	w_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1]} = 0.48$
- $P(Y = 1) = \frac{\exp[0.1]}{1+\exp[0.1]} = .52$
- Bias w_0 encodes the prior probability of a class

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
"viagra"	w_1	2.0
"mother"	w_2	-1.0
"work"	w_3	-0.5
"nigeria"	w_4	3.0

- What does $Y = 1$ mean?

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1-1.0+3.0]} = 0.13$
- $P(Y = 1) = \frac{\exp[0.1-1.0+3.0]}{1+\exp[0.1-1.0+3.0]} = .87$
- Include bias, and sum the other weights

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
“viagra”	w_1	2.0
“mother”	w_2	-1.0
“work”	w_3	-0.5
“nigeria”	w_4	3.0

- What does $Y = 1$ mean?

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1-1.0+3.0]} = 0.13$
- $P(Y = 1) = \frac{\exp[0.1-1.0+3.0]}{1+\exp[0.1-1.0+3.0]} = .87$
- Include bias, and sum the other weights

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
"viagra"	w_1	2.0
"mother"	w_2	-1.0
"work"	w_3	-0.5
"nigeria"	w_4	3.0

- What does $Y = 1$ mean?

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y = 0) = \frac{1}{1+\exp\left[0.1-1.0+3.0\right]} =$ 0.13
- $P(Y = 1) = \frac{\exp\left[0.1-1.0+3.0\right]}{1+\exp\left[0.1-1.0+3.0\right]} = .87$
- Include bias, and sum the other weights

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
“viagra”	w_1	2.0
“mother”	w_2	-1.0
“work”	w_3	-0.5
“nigeria”	w_4	3.0

- What does $Y = 1$ mean?

Example 3

$X = \{\text{Mother}, \text{Work}, \text{Viagra}, \text{Mother}\}$

- $P(Y=0) = \frac{1}{1+\exp\left[0.1-1.0-0.5+2.0-1.0\right]} = 0.13$
- $P(Y=1) = \frac{\exp\left[0.1-1.0-0.5+2.0-1.0\right]}{1+\exp\left[0.1-1.0-0.5+2.0-1.0\right]} = .87$
- Multiply feature presence by weight

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
"viagra"	w_1	2.0
"mother"	w_2	-1.0
"work"	w_3	-0.5
"nigeria"	w_4	3.0

- What does $Y = 1$ mean?

Example 3

$X = \{\text{Mother}, \text{Work}, \text{Viagra}, \text{Mother}\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0-0.5+2.0-1.0]} = 0.13$
- $P(Y=1) = \frac{\exp[0.1-1.0-0.5+2.0-1.0]}{1+\exp[0.1-1.0-0.5+2.0-1.0]} = .87$
- Multiply feature presence by weight

Logistic Regression Example

feature	symbol	weight
bias	w_0	0.1
"viagra"	w_1	2.0
"mother"	w_2	-1.0
"work"	w_3	-0.5
"nigeria"	w_4	3.0

- What does $Y = 1$ mean?

Example 3

$X = \{\text{Mother}, \text{Work}, \text{Viagra}, \text{Mother}\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0-0.5+2.0-1.0]} = 0.13$
- $P(Y=1) = \frac{\exp[0.1-1.0-0.5+2.0-1.0]}{1+\exp[0.1-1.0-0.5+2.0-1.0]} = .87$
- Multiply feature presence by weight

Equivalence of Naïve Bayes and Logistic Regression

Consider Naïve Bayes and logistic regression with two classes: (+) and (-).

Naïve Bayes

$$\hat{P}(c_+)\prod_i \hat{P}(w_i|c_+)$$

$$\hat{P}(c_-)\prod_i \hat{P}(w_i|c_-)$$

Logistic Regression

$$\sigma\left(-w_0 - \sum_i w_i X_i\right) = \frac{1}{1+\exp\left(w_0 + \sum_i w_i X_i\right)}$$

$$1-\sigma\left(-w_0 - \sum_i w_i X_i\right) = \frac{\exp\left(w_0 + \sum_i w_i X_i\right)}{1+\exp\left(w_0 + \sum_i w_i X_i\right)}$$

- These are actually the same if $w_0 = \sigma\left(\log\left(\frac{p(c_+)}{1-p(c_+)}\right) + \sum_j \log\left(\frac{1-P(w_j|c_+)}{1-P(w_j|c_-)}\right)\right)$
- and $w_j = \log\left(\frac{P(w_j|c_+)(1-P(w_J|c_-))}{P(w_j|c_-)(1-P(w_J|c_+))}\right)$

Contrasting Naïve Bayes and Logistic Regression

- Naïve Bayes easier
- Naïve Bayes better on smaller datasets
- Logistic regression better on medium-sized datasets
- On huge datasets, it doesn't really matter (data always wins)
    - Optional reading by Ng and Jordan has proofs and experiments
- Logistic regression allows arbitrary features (this is why naïve Bayes not in Rattle)

Contrasting Naïve Bayes and Logistic Regression

- Naïve Bayes easier
- Naïve Bayes better on smaller datasets
- Logistic regression better on medium-sized datasets
- On huge datasets, it doesn't really matter (data always wins)
  - Optional reading by Ng and Jordan has proofs and experiments
- Logistic regression allows arbitrary features (this is why naïve Bayes not in Rattle)
- Don't need to memorize (or work through) previous slide—just understand that naïve Bayes is a special case of logistic regression

Outline

Next time . . .

- More classification
  - State-of-the-art models
  - Interpretable models
  - Not the same thing!
- What does it mean to have a good classifier?
- Running all these classifiers in Rattle