GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 8: BAYESIAN INFERENCE

LAST TIME:

- LOGISTIC REGRESSION

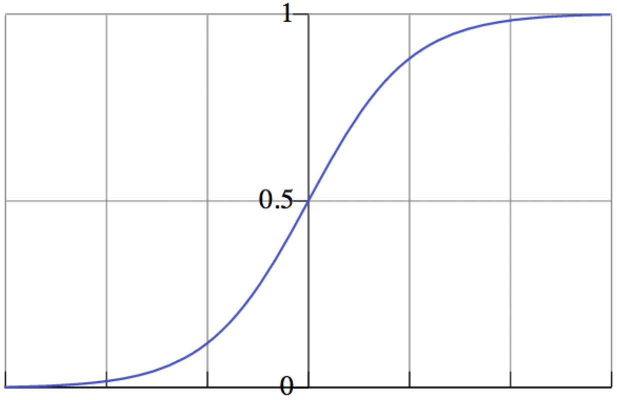
QUESTIONS?

I. REVIEW LOGISTIC REGRESSION
II. PROBABILITY
III. BAYESIAN INFERENCE

I. LOGISTIC REGRESSION

The logistic function*:*

$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$



The **logit function** *is an important transformation of the logistic function. Notice that it returns the linear model!*

$$g(x) = ln(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$$

I. INTRO TO PROBABILITY

Q: What is a **probability***?*

Q: What is a **probability***?*

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

Q: What is a **probability***?*

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

The probability of event A *is denoted* $P(A)$.

Q: What is the set of all possible events called?

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω. *Event* A *is a member of the sample space, as is every other event.*

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω. *Event* A *is a member of the sample space, as is every other event.*

The probability of the sample space $P(\Omega)$ *is 1.*

Q: What is a probability distribution?

A: A function that assigns probability to each event in the sample space.

Q: What is a probability distribution?
A: A function that assigns probability to each event in the sample space.

A distribution can be discrete or continuous
Ex:
Discrete – Uniform distribution

$$X \sim \{1, \dots, N\} \quad - \quad P(X = x) = 1/N$$

Q: What is a probability distribution?

A: A function that assigns probability to each event in the sample space.

A distribution can be discrete or continuous

Ex:

Continuous – Normal distribution – N(u, o)

$X \sim N(0, 1)$ - $P(X = x) = 0$

Q: What is expected value?

A: It is the average value of a random variable – one that represents the most common value

Q: What is expected value?

A: It is the average value of a random variable – one that represents the most common value

For discrete distributions

$$E(X) = \Sigma\ x * p(x)$$

For continuous distributions

$$E(X) = \text{integral}\ (\ x * p(x)\ )$$

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

1) *Linda is a bank teller.*
2) *Linda is a bank teller and active in the feminist movement.*

Q: Consider two events A & B. *How can we characterize the* **intersection** *of these events?*

Q: Consider two events A & B*. How can we characterize the* **intersection** *of these events?*

A: With the **joint probability** *of* A *and* B*, written* $P(AB)$.

Q: Suppose event B *has occurred. What quantity represents the probability of* A **given** *this information about* B*?*

Q: Suppose event B *has occurred. What quantity represents the probability of* A **given** *this information about* B*?*

A: The intersection of A & B *divided by region* B.

Q: Suppose event B *has occurred. What quantity represents the probability of* A **given** *this information about* B*?*

A: The intersection of A & B *divided by region* B.

Q: Suppose event B *has occurred. What quantity represents the probability of* A **given** *this information about* B*?*

A: The intersection of A & B *divided by region* B.

This is called the **conditional probability** *of* A *given* B*, written* $P(A|B) = P(AB) \ / \ P(B)$.

Q: Suppose event B *has occurred. What quantity represents the probability of* A **given** *this information about* B*?*

A: The intersection of A & B *divided by region* B.

This is called the **conditional probability** *of* A *given* B, *written* P(A|B) = P(AB) / P(B).

Notice, with this we can also write P(AB) = P(A|B) * P(B).

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

1) *Linda is a bank teller.*
2) *Linda is a bank teller and active in the feminist movement.*

Q: What does it mean for two events to be **independent***?*

Q: What does it mean for two events to be **independent***?*

A: Information about one does not affect the probability of the other.

Q: What does it mean for two events to be **independent***?*

A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

Q: What does it mean for two events to be **independent***?*

A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

Using the definition of the conditional probability, we can also write:

$$P(A|B) = P(AB) \ / \ P(B) = P(A) \ \rightarrow \qquad P(AB) = P(A) * P(B)$$

II. BAYESIAN INFERENCE

This result is called **Bayes' theorem***. Here it is again:*

$$P(A|B) = P(B|A) * P(A) \ / \ P(B)$$

This result is called **Bayes' theorem***. Here it is again:*

$$P(A|B) = P(B|A) * P(A) \;/\; P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

This result is called **Bayes' theorem**. *Here it is again:*

$$P(A|B) = P(B|A) * P(A) \ / \ P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.

This result is called **Bayes' theorem**. *Here it is again:*

P(A|B) = P(B|A) * P(A) / P(B)

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.

- It's a very powerful computational tool.

Each term in this relationship has a name, and each plays a distinct role in any probability calculation (including ours).

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **likelihood function**. *It represents the joint probability of observing features* $\{x_i\}$ *given that that record belongs to class* C.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **likelihood function**. *It represents the joint probability of observing features* $\{x_i\}$ *given that that record belongs to class* C.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

We can observe the value of the likelihood function from the training data.

This term is the **prior probability** *of* C*. It represents the probability of a record belonging to class* C *before the data is taken into account.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **prior probability** *of* C*. It represents the probability of a record belonging to class* C *before the data is taken into account.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The value of the prior is also observed from the data.

This term is the **normalization constant.** *It doesn't depend on* C, *and is generally ignored until the end of the computation.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Maximum likelihood estimator (MLE):

*What parameters **maximize** the likelihood function?*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Maximum a posteriori estimate (MAP):

What parameters ***maximize*** *the likelihood function* ***AND*** *prior?*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Problem:

We observe the following coin flips:

HTHH

What is $P(X = \text{Heads})$?

Problem:

We observe the following coin flips:

HTHH

What is P(X = Heads) ? 3/4, Why?

Problem:

We observe the following coin flips:

HTHHTHT

What is P(X = Heads) ?

Problem:

We observe the following coin flips:

HTHHTHT

What is $P(X = \text{Heads})$? 4/7, Why?

We observe the following coin flips:
HTHHTHT

Maximum likelihood estimator (MLE):
*What parameters **maximize** the likelihood function?*
Let P(X = Heads) = q, and write Bayes Theorem

P(q | observations) = P (observations | q) * P (q) / constant

Maximum likelihood estimator (MLE):

*What parameters **maximize** the likelihood function?*

Let P(X = Heads) = q, and write Bayes Theorem

P(q | observations) = P (observations | q) * P (q) / constant

P(observations | q) = ?

P(q) = ?

Maximum likelihood estimator (MLE):

What parameters ***maximize*** *the likelihood function?*

Let P(X = Heads) = q, and write Bayes Theorem

P(q | observations) = P (observations | q) * P (q) / constant

P(observations | q) = Binomial Distribution

P(q) = ????

Binomial Distribution:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

P (HTHHTHT | q) = P (X = 4, n = 7) =

= (7 choose 4) * q^4 * (1-q) ^ 3

Binomial Distribution:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

P (HTHHTHT | q) = P (X = 4, n = 7) =

= (7 choose 4) * q^4 * (1-q) ^ 3

After optimizing, the **MLE is 4/7**

A prior distribution is known as **conjugate prior** if its from the same family as the posterior for a certain likelihood function

For the binomial distribution, the conjugate prior is the **Beta distribution**

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The **MAP estimate** is the value that maximizes both the likelihood function and prior – the product of the two.

In the coin flip setting is the value that optimizes

$$P(HTHHTHT \mid q) * P(q)$$

The **MAP estimate** is the value that maximizes both the likelihood function and prior - the product of the two.

In the coin flip setting is the value that optimizes

P (HTHHTHT | q) * P(q)

= (7 choose 4) q ^ 4 * (1 -q) ^3 * q^(a-1) * (1-a) ^(b-1)

The **MAP estimate** is the value that maximizes both the likelihood function and prior - the product of the two.

In the coin flip setting is the value that optimizes

P (HTHHTHT | q) * P(q)

= (7 choose 4) q ^ 4 * (1 -q) ^3 * q^(a-1) * (1-a) ^(b-1)

= q^(4 + a -1) * (1-q)^ (3 + b - 1)

The **MAP estimate** is the value that maximizes both the likelihood function and prior – the product of the two.

In the coin flip setting is the value that optimizes

P (HTHHTHT | q) * P(q)

= (7 choose 4) q ^ 4 * (1 -q) ^3 * q^(a–1) * (1–a) ^(b–1)

= q^(4 + a -1) * (1–q)^ (3 + b – 1)

After optimizing, the **MAP is (4 + a -1) / (7 + a + b – 2)**

Why do you care?

Why do you care?

Many problems are binary and are estimated using counts...

Why do you care?

Many problems are binary and are estimated using counts...

Ex. 1:
Sample 100 people and ask if they support a politician?

Why do you care?

Many problems are binary and are estimated using counts...

Ex. 1:
Sample 100 people and ask if they support a politician?
23 say Yes - Is the correct prediction 23/100?
What's the prior?

Ex. 2:

Need to choose between multiple categories to present (for ads, products, news).

Ex. 2:

Need to choose between multiple categories to present (for ads, products, news).

You can compute response % for each category

Ex. 2:

Need to choose between multiple categories to present (for ads, products, news).

You can compute response % for each category

But each should have a unique prior – **unique psuedo counts**

Suppose we have a dataset with features $x_1, \ldots, x_n$ *and a class label* c. *What can we say about classification using Bayes' theorem?*

Suppose we have a dataset with features $x_1, \ldots, x_n$ *and a class label* C*. What can we say about classification using Bayes' theorem?*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Bayes' theorem can help us to determine the probability of a record belonging to a class, given the data we observe.

source: Data Analysis with Open Source Tools, by Philipp K. Janert. O'Reilly Media, 2011.

The idea of Bayesian inference, then, is to **update** *our beliefs about the distribution of C using the data ("evidence") at our disposal.*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Then we can use the posterior for prediction.

Q: What piece of the puzzle we've seen so far looks like it could intractably difficult in practice?

Remember the likelihood function?

$$P(\{x_i\}|C) = P(\{x_1, x_2, \ldots, x_n\})|C)$$

Remember the likelihood function?

$$P(\{x_i\}|C) = P(\{x_1, x_2, \ldots, x_n\})|C)$$

Observing this exactly would require us to have enough data for every possible combination of features to make a reasonable estimate.

Q: What piece of the puzzle we've seen so far looks like it could intractably difficult in practice?

A: Estimating the full likelihood function.

Q: So what can we do about it?

Q: So what can we do about it?

A: Make a simplifying assumption. In particular, we assume that the features x_i *are conditionally independent from each other:*

Q: So what can we do about it?

A: Make a simplifying assumption. In particular, we assume that the features x_i are conditionally independent from each other:

$$P(\{x_i\}|C) = P(x_1, x_2, \ldots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C)$$

Q: So what can we do about it?

A: Make a simplifying assumption. In particular, we assume that the features x_i *are conditionally independent from each other:*

$$P(\{x_i\}|C) = P(x_1, x_2, \ldots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C)$$

This "naïve" assumption simplifies the likelihood function to make it tractable.

$$P(\{x_i\}|C) = P(x_1, x_2, \ldots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C)$$

Q: Given that we can compute this value, what do we do with it?

$$P(\{x_i\}|C) = P(x_1, x_2, \ldots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C)$$

Q: Given that we can compute this value, what do we do with it?

A: In our training phase, we 'learn' the probability of seeing our training examples under each class.

$$P(\{x_i\}|C) = P(x_1, x_2, \ldots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C)$$

Q: Given that we can compute this value, what do we do with it?

A: In our training phase, we 'learn' the probability of seeing our training examples under each class.

Then we use Bayes Theorem to compute P(class | inputs)

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Maximum a posteriori estimate (MAP):

What ***LABEL maximizes*** *the likelihood function* ***AND*** *prior?*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Example: Text Classification

Does this news article talk about politics?

Training Set: Collection of New Articles

Example: Text Classification

Does this news article talk about politics?

Training Set: Collection of New Articles

Article 1: The computer contractor who exposed....
Article 2: The parents of a missing U.S. journalist in Syria...

Q: What are my features?

Q: What are my features?

A: The text in the documents.

Q: What are my features?

A: The text in the documents.

Q: How to I represent them?

Q: What are my features?

A: The text in the documents.

Q: How to I represent them?

A: Binary occurrence? Word counts?

the,	*computer,*	*contractor,*	*exposed,*	*parents,*	*missing,*	*Syria,*	*U.S.*
1	*1*	*1*	*1*	*0*	*0*	*0*	*0*
1	*0*	*0*	*0*	*1*	*1*	*1*	*1*

computer,	contractor,	exposed,	parents,	missing,	Syria,	U.S.
1	1	1	0	0	0	0
0	0	0	1	1	1	1

We can make some alterations

1) Drop stop words (commonly occurring words that don't have meaning)

computer,	contractor,	exposed,	parents,	missing,	Syria,	U.S.,	**POL**
1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	1

Our goal is to compute compute P (POL = T | words in the text)

We need to ***learn*** *P(word | POL) i.e. P (Syria | POL)*

computer,	contractor,	exposed,	parents,	missing,	Syria,	U.S.,	**POL**
1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	1

Once we've learned P(computer | POL), P(U.S. | POL) on our training set, we want to label our test set

computer,	contractor,	exposed,	parents,	missing,	Syria,	U.S.,	**POL**
1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	1

The correct label, POL = True or POL = False is the one that maximize our posterior.

computer,	*contractor,*	*exposed,*	*parents,*	*missing,*	*Syria,*	*U.S.,*	***POL***
1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	1

Compute probability in each class:

$$P(POL = T \mid \{x\}) = P(\{x\} \mid POL = T) * P(POL=T)$$

$$P(POL = F \mid \{x\}) = P(\{x\} \mid POL = F) * P(POL=F)$$

computer,	contractor,	exposed,	parents,	missing,	Syria,	U.S.,	**POL**
1	1	1	0	0	0	0	0
0	0	0	1	1	1	1	1

Article 2: The parents of a missing U.S. journalist in Syria...

$$P(POL = T \mid \{x\}) = P(\{x\} \mid POL = T) * P(POL=T)$$

$$= P(Syria \mid POL=T) * P(journalist \mid POL=T) * P(parents \mid POL=T) \ldots$$

$$* P(POL=T)$$