

# INTRO to DATA SCIENCE

## LECTURE 8: BAYESIAN INFERENCE

## **LAST TIME:**

- PROBABILITY**
- LOGISTIC REGRESSION**

## **QUESTIONS?**

**I. REVIEW LOGISTIC REGRESSION**

**II. BAYESIAN INFERENCE**

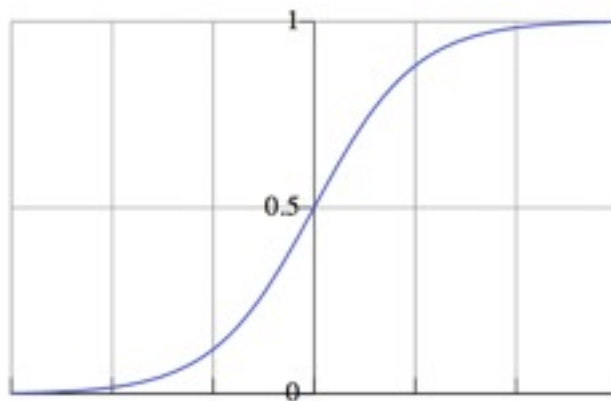
**EXERCISES:**

**III. IMPLEMENTING A SPAM FILTER**

# 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!*

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

# II. BAYESIAN INFERENCE

**Bayes' theorem.** *Here it is again:*

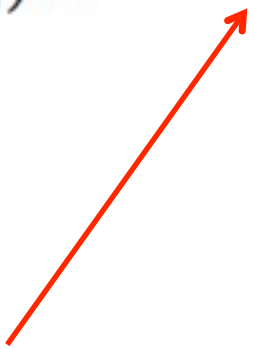
$$P(A|B) = P(B|A) * P(A) / P(B)$$

*Some facts:*

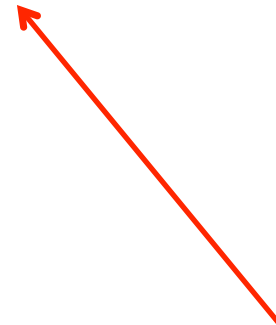
- This is a simple algebraic relationship using elementary definitions.*
- It's interesting because it's kind of a "wormhole" between two different "interpretations" of probability.*
- It's a very powerful computational tool.*



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

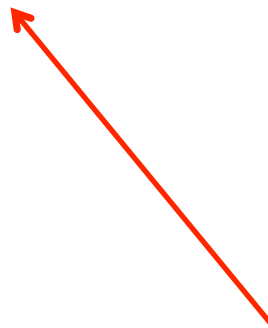
$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


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

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


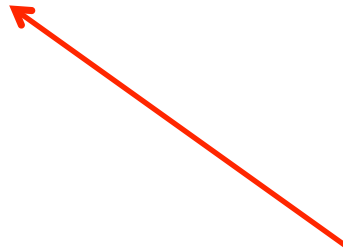
*This term is the **normalization constant**. It doesn't depend on  $C$ , and is generally ignored until the end of the computation.*

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



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

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



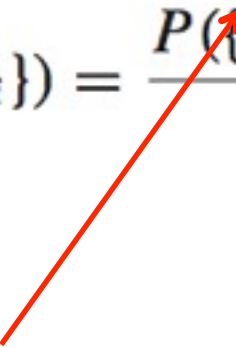
*This term is the **posterior probability** of  $C$ . It represents the probability of a record belonging to class  $C$  after the data is taken into account.*

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

*The goal of any Bayesian computation is to find (“learn”) the posterior distribution of a particular variable.*

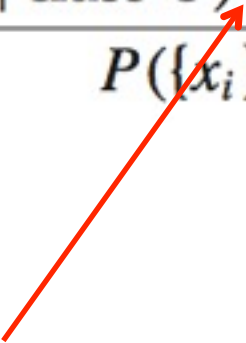
*Maximum likelihood estimator (MLE):*

*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  $P(X = \text{Heads})$  ?  $3/4$ , Why?**

**Problem:**

**We observe the following coin flips:**

**HTHHTHT**

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

**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  $P(X = \text{Heads}) = q$ , and write Bayes Theorem

$$P(q \mid \text{observations}) = P(\text{observations} \mid q) * P(q) / \text{constant}$$

*Maximum likelihood estimator (MLE):*

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

Let  $P(X = \text{Heads}) = q$ , and write Bayes Theorem

$$P(q \mid \text{observations}) = P(\text{observations} \mid q) * P(q) / \text{constant}$$

$$P(\text{observations} \mid q) = ?$$

$$P(q) = ?$$

*Maximum likelihood estimator (MLE):*

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

Let  $P(X = \text{Heads}) = q$ , and write Bayes Theorem

$$P(q \mid \text{observations}) = P(\text{observations} \mid q) * P(q) / \text{constant}$$

$P(\text{observations} \mid q) = \text{Binomial Distribution}$

$P(q) = \text{????}$

## Binomial Distribution:

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

$$\begin{aligned} P(\text{HTHHTHT} \mid q) &= P(X = 4, n = 7) = \\ &= (7 \text{ choose } 4) * q^4 * (1-q)^3 \end{aligned}$$

## Binomial Distribution:

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

$$\begin{aligned} P(\text{HTHHTHT} \mid q) &= P(X = 4, n = 7) = \\ &= \binom{7}{4} * q^4 * (1-q)^3 \end{aligned}$$

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**

$$\begin{aligned} &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \\ &= \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \end{aligned}$$

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  
$$P(\text{HTHHTHT} \mid q) * P(q)$$

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

$$\begin{aligned} & P(\text{HTHHTHT} \mid q) * P(q) \\ &= \binom{7}{4} q^4 * (1-q)^3 * q^{(a-1)} * (1-a)^{(b-1)} \end{aligned}$$

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

$$\begin{aligned} & P(\text{HTHHTHT} \mid q) * P(q) \\ &= \binom{7}{4} q^4 * (1-q)^3 * q^{(a-1)} * (1-q)^{(b-1)} \\ &= q^{(4+a-1)} * (1-q)^{(3+b-1)} \end{aligned}$$

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

$$\begin{aligned} & P(\text{HTHHTHT} \mid q) * P(q) \\ &= \binom{7}{4} q^4 * (1-q)^3 * q^{(a-1)} * (1-q)^{(b-1)} \\ &= q^{(4+a-1)} * (1-q)^{(3+b-1)} \end{aligned}$$

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  $x_1, \dots, x_n$  and a class label  $c$ .  
What can we say about classification using Bayes' theorem?*

*Suppose we have a dataset with features  $x_1, \dots, x_n$  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?*

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

*Remember the likelihood function?*

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

*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:*

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

*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:*

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

*This “naïve” assumption simplifies the likelihood function to make it tractable.*

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

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



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

*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.*

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

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

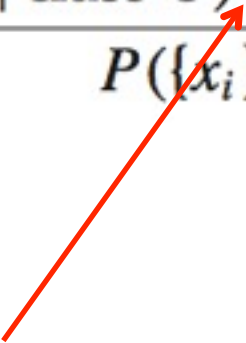
*A: In our training phase, we ‘learn’ the probability of seeing our training examples under each class.*

*Then we use Bayes Theorem to compute  $P(\text{class} | \text{inputs})$*

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

*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.*

<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

*computer, contractor, exposed, parents, missing, Syria, U.S.*

<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

*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***

<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

*Our goal is to compute  $P ( POL = T \mid \text{words in the text} )$*

*We need to **learn**  $P( \text{word} \mid POL )$  i.e.  $P ( Syria \mid POL )$*

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

<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

*Once we've learned  $P(\text{computer} \mid \text{POL})$ ,  $P(\text{U.S.} \mid \text{POL})$  on our training set, we want to label our test set*

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

<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

*The correct label,  $POL = \text{True}$  or  $POL = \text{False}$  is the one that maximize our posterior.*

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

<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

*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***

<i>1</i>	<i>1</i>	<i>1</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
<i>0</i>	<i>0</i>	<i>0</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>

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

$$\begin{aligned}
 P ( POL = T \mid \{x\} ) &= P ( \{x\} \mid POL = T ) * P(POL=T) \\
 &= P(Syria \mid POL=T) * P(journalist \mid POL=T) * P(parents \mid POL=T) ... \\
 &\quad * P( POL=T)
 \end{aligned}$$