GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 7: PROBABILITY AND NAIVE BAYES CLASSIFICATION

Paul Burkard

11/16/2015

LAST TIME:

- CLASSIFICATION
- K-NEAREST NEIGHBORS CLASSIFICATION

QUESTIONS?

I INTRO TO PROBABILITY

II. BAYESIAN INFERENCE

III. NAIVE BAYES CLASSIFICATION

HANDS-ON: NAIVE BAYES CLASSIFICATION

I. INTRO TO PROBABILITY

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?

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.

A distribution can be **discrete** *or* **continuous**

Ex: Discrete – Uniform distribution

$$X \sim \{1, \ldots, N\} \longrightarrow P(X = x) = 1/N$$

Continuous – Normal distribution – $N(\mu, \sigma^2)$

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

Discrete Probability Distributions:

- *These are* **probability mass functions**
- *Each value P(X=x) represents the probability of observing a given value x for variable X*

$$P(\Omega) = \sum_{X=x} P(X = x) = 1$$

Continuous Probability Distributions:

- *These are* **probability density functions (PDF)**
- *Each value P(X=x) represents the* **relative probability** *of observing a given value x for variable X*

$$P(x_0 < x < x_1) = \int_{x_0}^{x_1} P(x)\,dx$$

$$P(\Omega) = \int_{-\infty}^{+\infty} P(x)\,dx = 1$$

Q: What is a random variable?

A: Essentially, a measurable whose possible values have a probability distribution

Values of these are the "Events" for which we're looking to measure the probabilities

Q: What is expected value?

A: It is the average value of a random variable

For discrete distributions

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

For continuous distributions

$$E(X) = \int(x * p(x))dx$$

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* $P(A, B)$*.*

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 $P(A|B) = P(A \cap B) \,/\, P(B)$.*

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:

$$P(A|B) = P(A \cap B) \,/\, P(B) = P(A) \rightarrow P(A \cap B) = P(A) * P(B)$$

II. BAYESIAN INFERENCE

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

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

Some facts:

- This is a simple algebraic relationship using elementary definitions.

- 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). Here's how it might look in the context of classification.

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

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?*

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

III. NAIVE BAYES CLASSIFICATION

	continuous	categorical
supervised	regression	classification
unsupervised	dim reduction	clustering

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.*

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

The goal of any Bayesian computation is to find ("learn") the posterior distribution of a particular variable given new **evidence**.

Problem:

We observe the following coin flips:

HTHH

What is P(X = 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) = ?

Binomial Distribution:

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

P (4 heads, 3 tails | 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 (4 heads, 3 tails | q) = P (X = 4, n = 7)
= (7 choose 4) * q^4 * $(1-q)^3$

Optimize w.r.t. q —> **MLE:** q = 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

$$P(4H, 3T \mid q) * P(q)$$

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

In the coin flip setting with a Beta distribution it's the value that optimizes:

$$P(4H, 3T \mid q) * P(q)$$

$$= (7 \text{ choose } 4)\, q^4(1-q)^3 q^{(\alpha-1)}(1-q)^{(\beta-1)}$$

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?

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.

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 be 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)

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 do 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*

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

We can make some alterations

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

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

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

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

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

The predicted label, POL = True or
POL = False is the one that maximizes our posterior.

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

Compute probability in each class:

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

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

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

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

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

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

HANDS-ON: NAIVE BAYES