Metis

MR. BAYES AND HIS NAÏVETÉ

Paul Burkard
10/12/2016

I. BAYESIAN INFERENCE

II. BAYESIAN CLASSIFICATION

III. NAIVE BAYES CLASSIFICATION

HANDS-ON: NAIVE BAYES CLASSIFICATION

I. BAYESIAN INFERENCE

BAYES' THEOREM

Back to **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.

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

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

Consider a situation in which we have some set of beliefs:

- *Prior probability of an event (or probability distribution for a RV)*
- *Experiment and observe outcomes related to the event or RV*
- *How do we update our beliefs?*
- *Answer: Bayes' Rule!*
  - *This is "Bayesian Inference"!*

SOME TERMINOLOGY

Each term in this relationship has a name, and each plays a distinct role in any probability calculation (including ours). Here's the formula reframed in terms of some common terminology:

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

THE LIKELIHOOD FUNCTION

This term is the **likelihood function**. *It represents the joint probability of observing some set of outcomes given another known outcome.*

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

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

THE PRIOR

This term is the **prior probability** *of A. It represents the probability of a set of outcomes before the data is taken into account.*

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

The value of the prior can also be observed from the data.

THE NORMALIZATION CONSTANT

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

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

This term is the **posterior probability** *of A. It represents our updated probability of A given our observations that B has occurred.*

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

The posterior is our updated beliefs from our prior, given the evidence observed.

Example 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 = 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) = p, and write Bayes Theorem

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

Maximum likelihood estimator (MLE):

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

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

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

P(observations | p) = ?

P(p) = ?

Binomial Distribution:

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

P (4 heads, 3 tails | p) = P (X = 4, n = 7)

= (7 choose 4) * p^4 * $(1-p)^3$

Binomial Distribution:

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

P (4 heads, 3 tails | p) = P (X = 4, n = 7)

= (7 choose 4) * p^4 * $(1\text{-}p)^3$

Optimize w.r.t. p —> **MLE:** p = 4/7

Maximum likelihood estimator (MLE):

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

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

MAP ESTIMATE

But what if we want to take our prior beliefs into account?
e.g. we believe a priori that most coins are fair!

This leads to the "MAP Estimate"

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Maximum a posteriori estimate (MAP):

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

In other words, what parameter(s) maximize the **posterior***?*

Bayes' rule works just fine for RVs too:

$$P(X|Y) = P(X) * P(Y|X) / P(Y)$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

- *Start with a* **prior distribution** *(beliefs) for X*
- *Update it to* **posterior distribution** *for X based on evidence from Y*

Nice Property:
If a posterior distribution will have the same class of distribution (Beta, Gaussian, etc) as the prior distribution.

Why is this nice?
We keep the same distribution, just update its “shape parameters” as we observe evidence!
e.g.: Update mean for Gaussian

CONJUGATE PRIORS

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

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

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:

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

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

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?

Bayesian Inference:

- *Start with prior (optional) beliefs about event or RV*
- *Observe evidence*
- *Update beliefs as to probability of event or distribution of RV of interest*

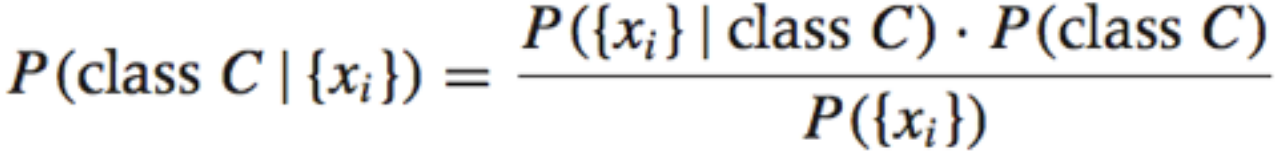
II. BAYESIAN CLASSIFICATION

TYPES OF LEARNING PROBLEMS

	continuous	categorical
supervised	regression	classification
unsupervised	dim reduction	clustering

BAYES RULE FOR CLASSIFICATION

Here's how Bayes' Theorem might look in the context of classification:



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

THE LIKELIHOOD FUNCTION

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

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

THE PRIOR

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 NORMALIZATION CONSTANT

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

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

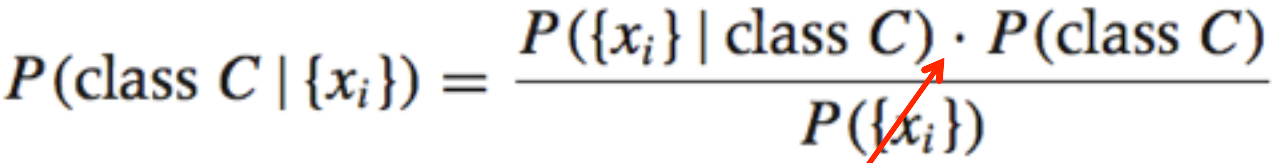
Maximum likelihood estimator (MLE):

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?*



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

THE POSTERIOR

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.*

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

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

III. NAIVE BAYES CLASSIFICATION

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?

NAÏVE BAYESIAN CLASSIFICATION

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?

NAÏVE BAYESIAN CLASSIFICATION

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

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?

NAÏVE BAYESIAN CLASSIFICATION

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?

NAÏVE BAYESIAN CLASSIFICATION

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

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*

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)

NAÏVE BAYESIAN CLASSIFICATION

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

NAÏVE BAYESIAN CLASSIFICATION

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.

NAÏVE BAYESIAN CLASSIFICATION

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:

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

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

NAÏVE BAYESIAN CLASSIFICATION

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: BAYESIAN COIN FLIPPING