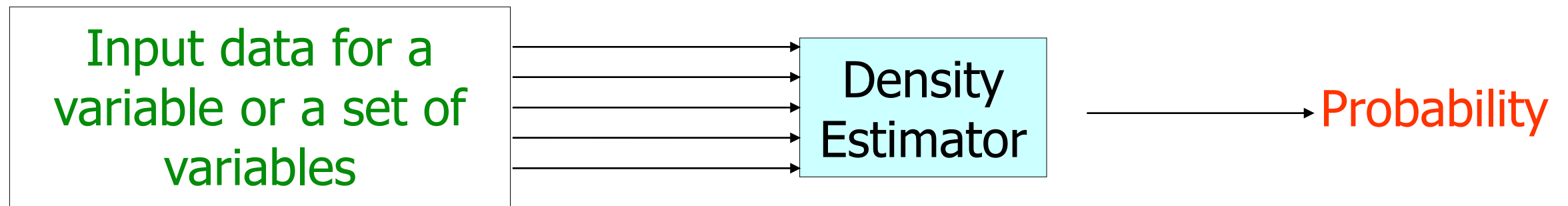


Density estimation

Density Estimation

- A Density Estimator learns a mapping from a set of attributes to a Probability



Density estimation

- Estimate the distribution (or conditional distribution) of a random variable
- Types of variables:
 - Binary
coin flip, alarm
 - Discrete
dice, car model year
 - Continuous
height, weight, temp.,

When do we need to estimate densities?

- Density estimators are critical ingredients in several of the ML algorithms we will discuss
- In some cases these are combined with other inference types for more involved algorithms (i.e. EM) while in others they are part of a more general process (learning in BNs and HMMs)

Density estimation

- Binary and discrete variables:

Easy: Just count!

- Continuous variables:

Harder (but just a bit): Fit
a model

Learning a density estimator for discrete variables

$$\hat{P}(x_i = u) = \frac{\text{\# records in which } x_i = u}{\text{total number of records}}$$

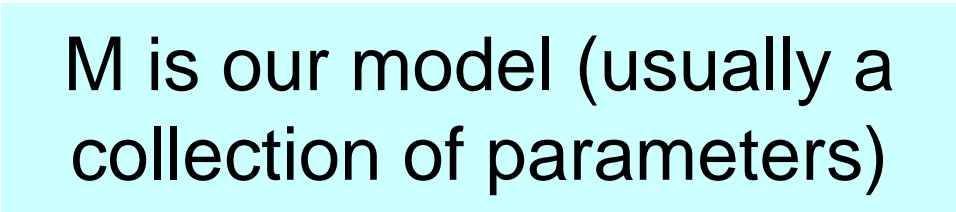
A trivial learning algorithm!

But why is this true?

Maximum Likelihood Principle

We can define the likelihood of the data given the model as follows:

$$\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \dots \wedge x_n \mid M) = \prod_{k=1}^n \hat{P}(x_k \mid M)$$



M is our model (usually a collection of parameters)

For example M is

- The probability of 'head' for a coin flip
- The probabilities of observing 1,2,3,4 and 5 for a dice
- etc.

Maximum Likelihood Principle

$$\hat{P}(\text{dataset} \mid M) = \hat{P}(x_1 \wedge x_2 \dots \wedge x_n \mid M) = \prod_{k=1}^n \hat{P}(x_k \mid M)$$

- Our goal is to determine the values for the parameters in M
- We can do this by maximizing the probability of generating the observed samples

- For example, let Θ be the probabilities for a coin flip

- Then

$$L(x_1, \dots, x_n \mid \Theta) = p(x_1 \mid \Theta) \dots p(x_n \mid \Theta)$$

- The observations (different flips) are assumed to be independent
- For such a coin flip with $P(H)=q$ the best assignment for Θ_h is

$$\text{argmax}_q = \#H/\#\text{samples}$$

- Why?

Maximum Likelihood Principle: Binary variables

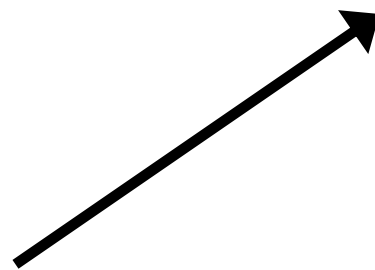
- For a binary random variable A with $P(A=1)=q$
 $\operatorname{argmax}_q = \#1/\#\text{samples}$

- Why?

Data likelihood: $P(D | M) = q^{n_1} (1 - q)^{n_2}$

We would like to find: $\operatorname{argmax}_q q^{n_1} (1 - q)^{n_2}$

Omitting terms that do not depend on q



Maximum Likelihood Principle

Data likelihood: $P(D | M) = q^{n_1} (1 - q)^{n_2}$

We would like to find: $\arg \max_q q^{n_1} (1 - q)^{n_2}$

$$\frac{\partial}{\partial q} q^{n_1} (1 - q)^{n_2} = n_1 q^{n_1 - 1} (1 - q)^{n_2} - q^{n_1} n_2 (1 - q)^{n_2 - 1}$$

$$\frac{\partial}{\partial q} = 0 \Rightarrow$$

$$n_1 q^{n_1 - 1} (1 - q)^{n_2} - q^{n_1} n_2 (1 - q)^{n_2 - 1} = 0 \Rightarrow$$

$$q^{n_1 - 1} (1 - q)^{n_2 - 1} (n_1 (1 - q) - q n_2) = 0 \Rightarrow$$

$$n_1 (1 - q) - q n_2 = 0 \Rightarrow$$

$$n_1 = n_1 q + n_2 q \Rightarrow$$

$$q = \frac{n_1}{n_1 + n_2}$$

Recall: Your first consulting job

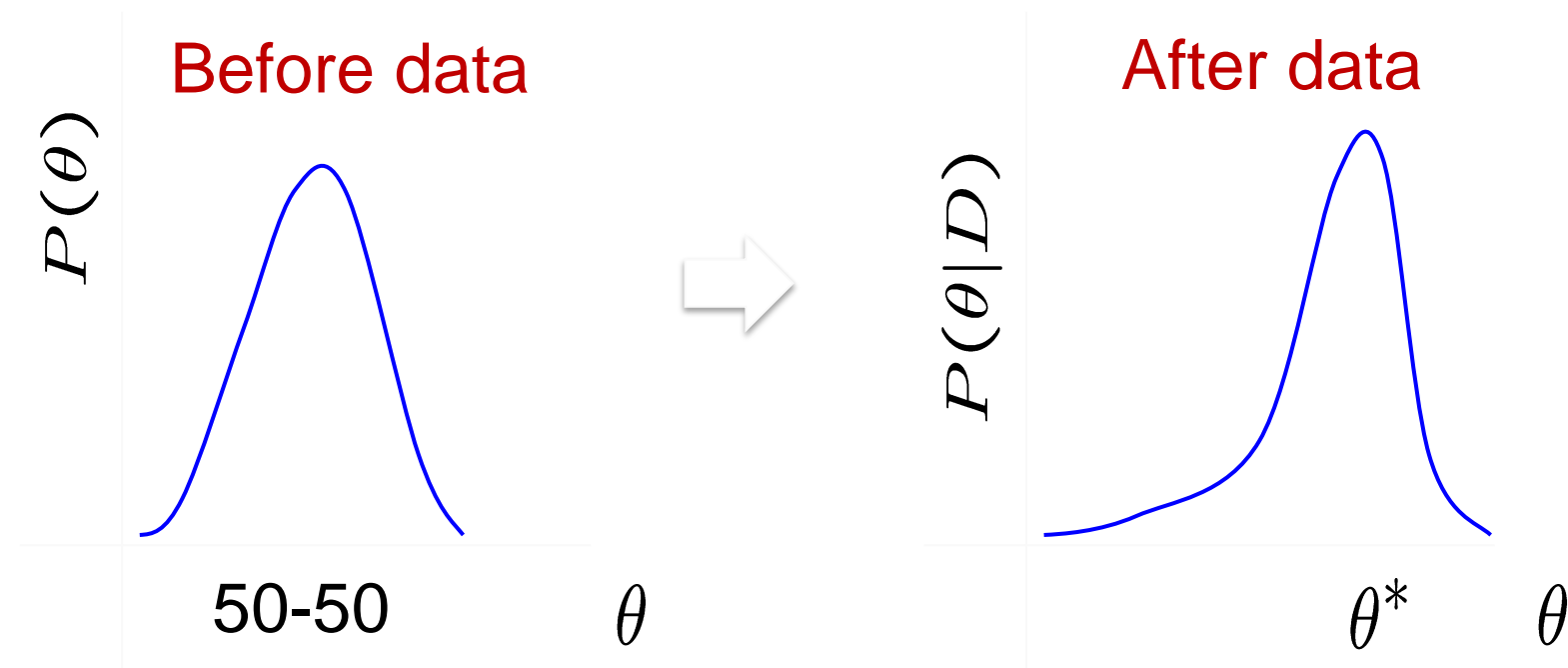
- A billionaire from the suburbs of Seattle asks you a question:
 - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
 - You say: Please flip it a few times:



- You say: The probability is: **3/5** because... frequency of heads in all flips
- **He says: But can I put money on this estimate?**
- You say: ummm.... Maybe not.
 - Not enough flips (less than sample complexity)

What about prior knowledge?

- Billionaire says: Wait, I know that the coin is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way...**
- Rather than estimating a single θ , we obtain a distribution over possible values of θ



Bayesian Learning

- Use Bayes rule:

$$P(\theta | \mathcal{D}) = \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)P(\theta)$$

posterior likelihood prior



Prior distribution

- From where do we get the prior?
 - Represents expert knowledge (philosophical approach)
 - Simple posterior form (engineer's approach)
- Uninformative priors:
 - Uniform distribution
- Conjugate priors:
 - Closed-form representation of posterior
 - $P(q)$ and $P(q|D)$ have the same algebraic form as a function of θ

Conjugate Prior

- $P(q)$ and $P(q|D)$ have the same form as a function of theta

Eg. 1 Coin flip problem

Likelihood given Bernoulli model:

$$P(\mathcal{D} | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

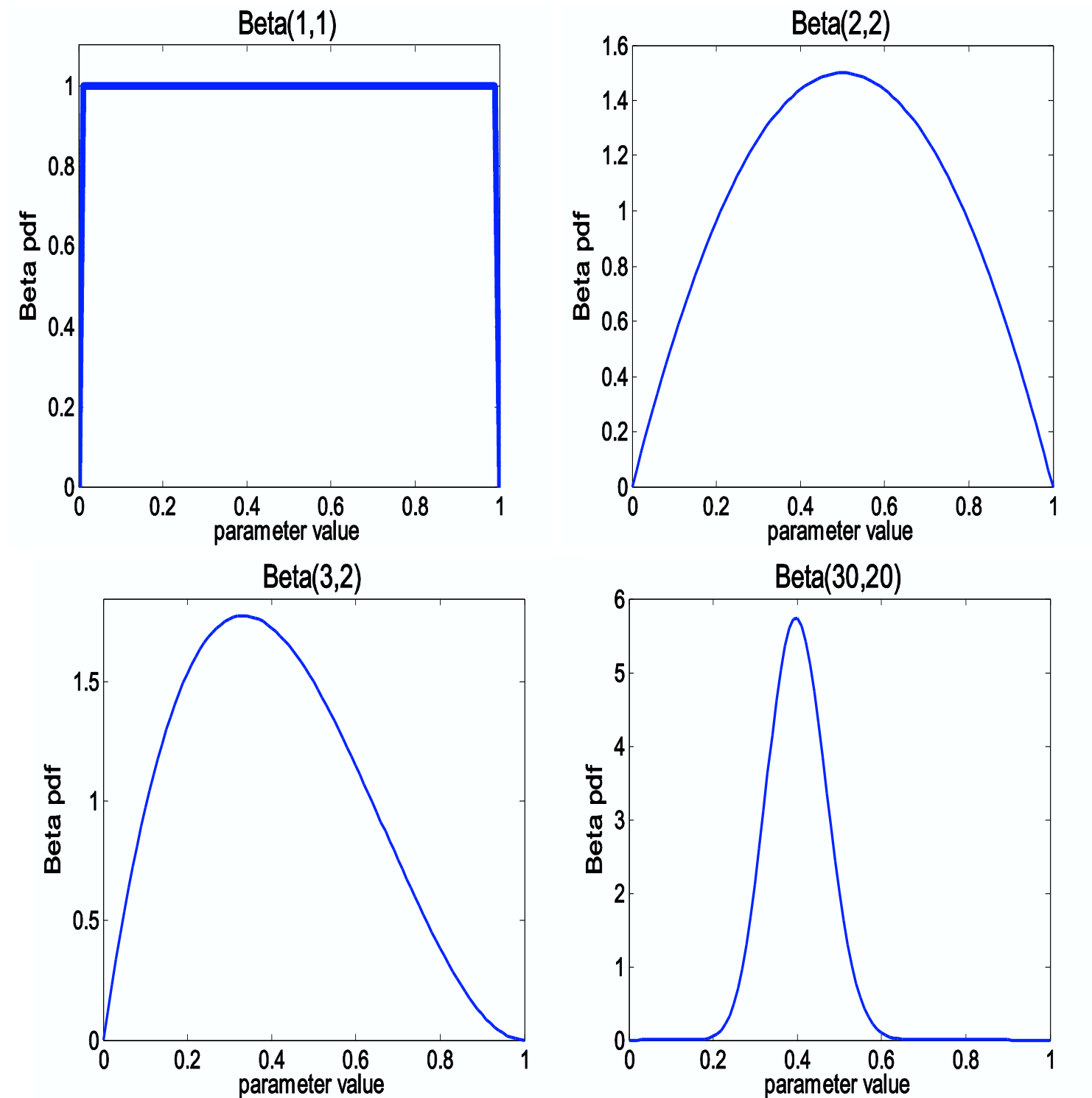
Then posterior is Beta distribution

$$P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



Beta distribution

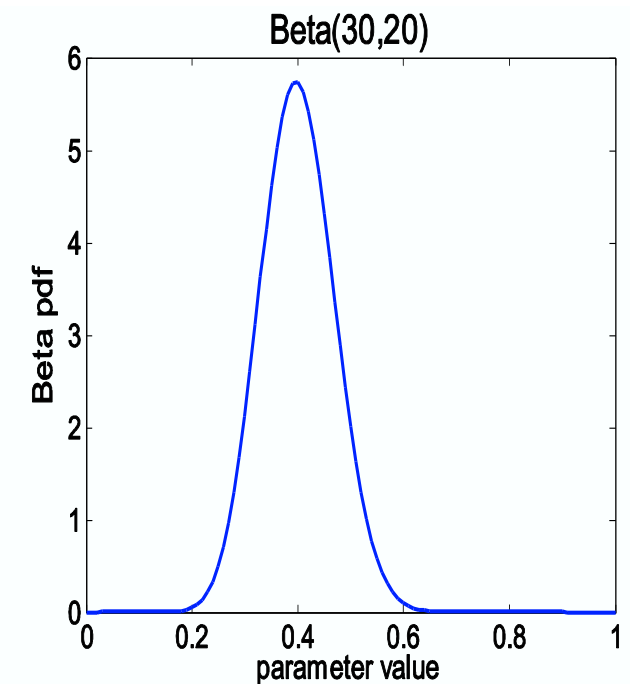
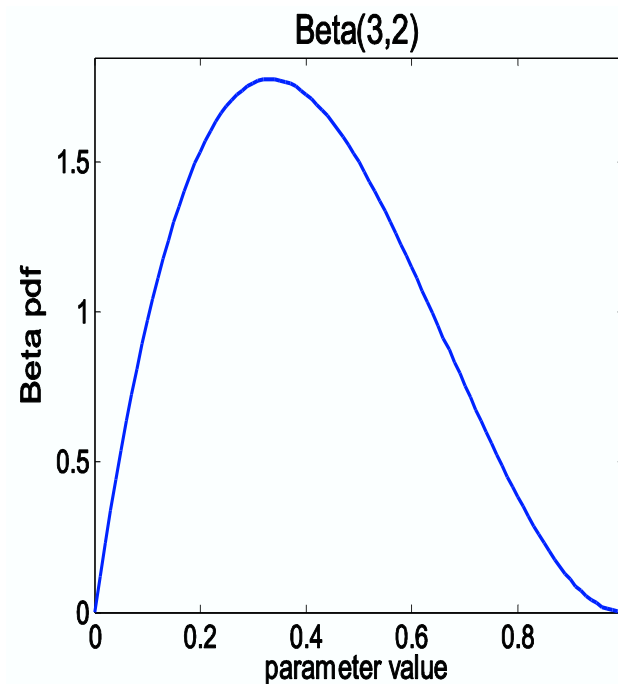
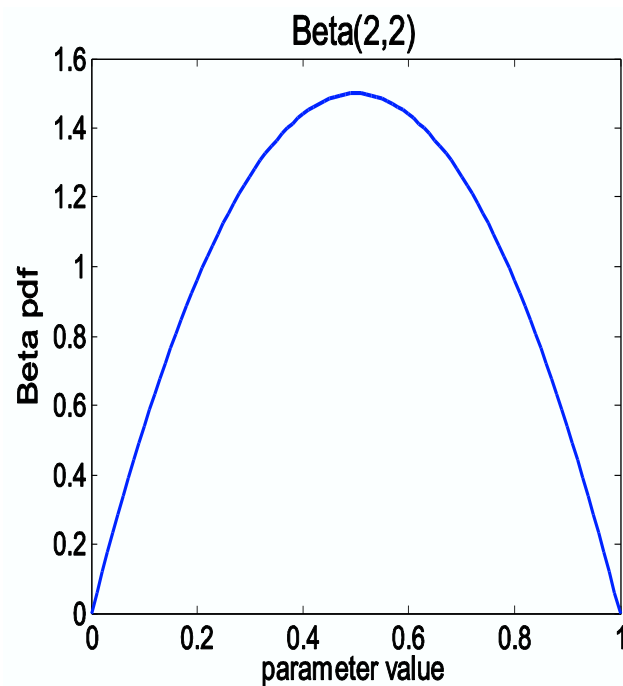
$Beta(\beta_H, \beta_T)$ More concentrated as values of β_H, β_T increase



Beta conjugate prior

$$P(\theta) \sim \text{Beta}(\beta_H, \beta_T)$$

$$P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$



As $n = \alpha_H + \alpha_T$
increases

As we get more samples, effect of prior is “washed out”

Conjugate Prior

The posterior $p(\theta | x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$

$$P(\mathcal{D}|\theta) = P(h, t|\theta) = \binom{n}{h} \theta^h (1 - \theta)^t = \binom{n}{h} \theta^h (1 - \theta)^{n-h}$$

$$P(\theta) = \frac{\theta^{\alpha-1} (1 - \theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$\begin{aligned} P(\theta|\mathcal{D}) &= \frac{P(\mathcal{D}|\theta)P(\theta)}{\int P(\mathcal{D}|\theta)P(\theta)d\theta} \\ &= \frac{\binom{n}{h} \theta^{h+\alpha-1} (1 - \theta)^{t+\beta-1} / B(\alpha, \beta)}{\int_{\theta=0}^1 \left(\binom{n}{h} \theta^{h+\alpha-1} (1 - \theta)^{t+\beta-1} / B(\alpha, \beta) \right) d\theta} \\ &= \frac{\theta^{h+\alpha-1} (1 - \theta)^{t+\beta-1}}{B(h + \alpha, t + \beta)} \end{aligned}$$

Conjugate Prior

- $P(\theta)$ and $P(\theta|D)$ have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is \sim Multinomial($\theta = \{\theta_1, \theta_2, \dots, \theta_k\}$)



$$P(\mathcal{D} | \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \dots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

Then poste

$$P(\theta) = \frac{\prod_{i=1}^k \theta_i^{\beta_i - 1}}{B(\beta_1, \dots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \dots, \beta_k)$$

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \dots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

Posterior Distribution

- The approach seen so far is what is known as a **Bayesian** approach
- Prior information encoded as a **distribution** over possible values of parameter
- Using the Bayes rule, you get an updated **posterior** distribution over parameters, which you provide with flourish to the Billionaire
- But the billionaire is not impressed
 - Distribution? I just asked for one number: is it $3/5$, $1/2$, what is it?
 - How do we go from a distribution over parameters, to a single estimate of the true parameters?

Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} P(\theta | D) \\ &= \arg \max_{\theta} P(D | \theta)P(\theta)\end{aligned}$$

MAP estimate of probability of head:

$$P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \quad \text{Mode of Beta distribution}$$

Naïve Bayes Classifier

- Bayes Classifier with additional “naïve” assumption:

- Features are independent given class:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

- More generally:

$$P(X_1 \dots X_d|Y) = \prod_{i=1}^d P(X_i|Y)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{bmatrix}$$

- If conditional independence assumption holds, NB is optimal classifier! But worse otherwise.

Conditional Independence

- X is **conditionally independent** of Y given Z:
probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- Equivalent to:

$$P(X, Y | Z) = P(X | Z)P(Y | Z)$$

- e.g., $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

Note: does NOT mean Thunder is independent of Rain

Conditional vs. Marginal Independence

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Conditional vs. Marginal Independence

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

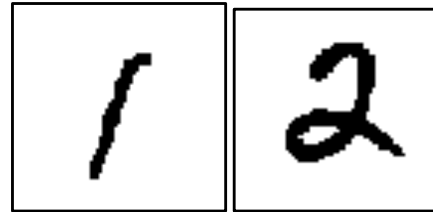
Finally another study pointed out that people wear coats when it rains...

Wearing coats is independent of accidents conditioned on the fact that it rained

Handwritten digit recognition (discrete features)

Training Data:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_d \end{bmatrix}$$



... n black-white (1/0)
images with
d pixels

Y

1

2

... n labels

How many parameters?

Class probability $P(Y = y) = p_y$ for all y **K-1 if K labels**

May not
hold

Class conditional distribution of features (using Naïve Bayes assumption)

$P(X_i = x_i | Y = y)$ – one probability value for each y, pixel i **Kd**

Linear instead of Exponential in d!

Naïve Bayes Classifier

- Bayes Classifier with additional “naïve” assumption:
 - Features are independent given class:

$$P(X_1 \dots X_d | Y) = \prod_{i=1}^d P(X_i | Y)$$

$$\begin{aligned} f_{NB}(\mathbf{x}) &= \arg \max_y P(x_1, \dots, x_d | y) P(y) \\ &= \arg \max_y \prod_{i=1}^d P(x_i | y) P(y) \end{aligned}$$

- Has fewer parameters, and hence requires fewer training data, even though assumption may be violated in practice

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$

- **Maximum Likelihood Estimates**

- For Class probability

$$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$

- For class conditional distribution

$$\frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

- NB Prediction for test data $X = (x_1, \dots, x_d)$

$$Y = \arg \max_y \hat{P}(y) \prod_{i=1}^d \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^n$ $X^{(j)} = (X_1^{(j)}, \dots, X_d^{(j)})$
- Maximum A Posteriori (MAP) Estimates – add m “virtual” datapts

Assume given some prior distribution (typically uniform):

$$Q(Y = b) \qquad Q(X_i = a, Y = b)$$

$$\hat{P}(X_i = a | Y = b) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = b\} + mQ(X_i = a, Y = b)}{\{\#j : Y^{(j)} = b\} + \underbrace{mQ(Y = b)}_{\text{\# virtual examples with } Y = b}}$$

virtual examples
with $Y = b$

Now, even if you never observe a class/feature posterior probability never zero.

Case Study: Text Classification

- Classify e-mails
 - $Y = \{\text{Spam, NotSpam}\}$
- Classify news articles
 - $Y = \{\text{what is the topic of the article?}\}$
- Classify webpages
 - $Y = \{\text{Student, professor, project, ...}\}$
- What about the features **X**?
 - The text!

Bag of words approach

the world of

TOTAL



all about the company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

► All About The Company

- Global Activities
- Corporate Structure
- TOTAL's Story
- Upstream Strategy
- Downstream Strategy
- Chemicals Strategy
- TOTAL Foundation
- Homepage



aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

Bag of words model

- Typical additional assumption – **Position in document doesn't matter**
 - “Bag of words” model – order of words on the page ignored
 - Sounds really silly, but often works very well!

in is lecture lecture next over person remember room
sitting the the the to to up wake when you

Bag of words model

- Typical additional assumption – **Position in document doesn't matter**
 - “Bag of words” model – order of words on the page ignored
 - Sounds really silly, but often works very well!

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

NB for Text Classification

- Features \mathbf{X} are the count of how many times each word in the vocabulary appears in document
- Probability table for $P(\mathbf{X}|Y)$ is huge!!!
- NB assumption helps a lot!!!
- Bag of words + Naïve Bayes assumption imply $P(\mathbf{X}|Y=y)$ is just the product of probability of each word, raised to its count, in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg \max_y P(y) \prod_{w=1}^W P(w|y)^{count_w}$$

NB with Bag of Words for text classification

- Learning phase:
 - Class Prior $P(Y)$: fraction of times topic Y appears in the collection of documents
 - $P(w|Y)$: fraction of times word w appears in documents with topic Y
- Test phase:
 - For each document
 - Use Bag of words + naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg \max_y P(y) \prod_{w=1}^W P(w|y)^{count_w}$$

Twenty news groups results

Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

Discriminative vs Generative Classifiers

Optimal Classifier:

$$\begin{aligned} f^*(x) &= \arg \max_{Y=y} P(Y = y | X = x) \\ &= \arg \max_{Y=y} P(X = x | Y = y) P(Y = y) \end{aligned}$$

Generative (Model based) approach: e.g. Naïve Bayes

- Assume some probability model for $P(Y)$ and $P(X|Y)$
- Estimate parameters of probability models from training data

Discriminative (Model free) approach: e.g. Logistic regression

Why not learn $P(Y|X)$ directly? Or better yet, why not learn the decision boundary directly?

- Assume some functional form for $P(Y|X)$ or for the decision boundary
- Estimate parameters of functional form directly from training data