Y-DATA
SCHOOL OF DATA SCIENCE

Naïve Bayes

Noa Lubin & Lior Sidi

Y-DATA
SCHOOL OF DATA SCIENCE

Agenda

- Motivation
- Theory
- Practical Example
- Naive Bayes Language Model
- Code
- Summary

Motivation

Y-DATA
SCHOOL OF DATA SCIENCE

Guess what’s in the box dogs or cats?

Problem: Guess what’s in the box dogs or cats?

we **cannot open the box**.

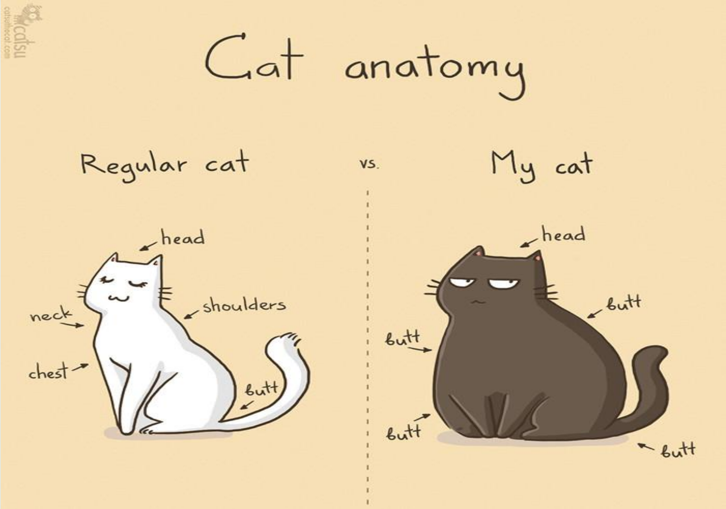
we can only **weight the box**.

We Know That

Dogs - 471M (56%)

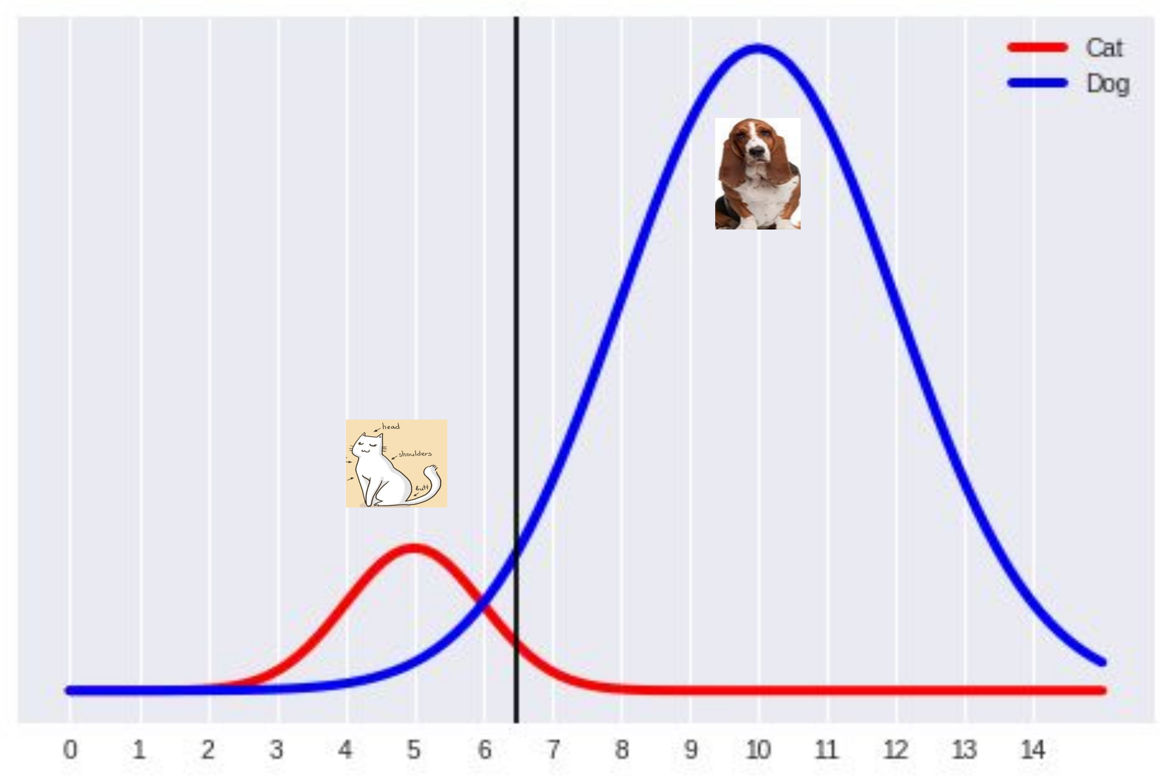
Cats - 373M (44%)

$p(\text{weight}|\text{Cat}) \sim N(5,1)$



$p(\text{weight}|\text{Dog}) \sim N(10,4)$

A box weighs 6.5kg, is it a Dog or a Cat?



Theory

Reminder:

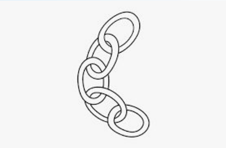
Independence assumption

Independent events:

E and F are *independent* if $P(EF) = P(E)P(F) \Leftrightarrow P(E|F) = P(E)$

Reminder:

Chain Rule



Conditional probability:

$$P(E|F) = \frac{P(EF)}{P(F)} \qquad \text{or} \qquad P(EF) = P(E|F)P(F)$$

Chaining rule:

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1 A_2 \ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2)\cdots P(A_n|A_1 A_2 \ldots A_{n-1})$$

Y-DATA
SCHOOL OF DATA SCIENCE

Reminder:

Bayes Rule

Bayes' formula:

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

Estimation

We are interested in:

- Classification - Ex: Decide the topic of a sentence.
- Regression - Ex: Weather
- Generation = Ex: Generate text

We’ll look at the generative behavior that causes something.

What we formally call: maximum likelihood estimation

Intuition Behind MLE

Let say we have a coin with bernoulli distribution for heads = ?

We can toss the coin 100 times. Let’s say that it landed on heads 68 times.

What do you think the probability for heads is?

Y-DATA
SCHOOL OF DATA SCIENCE

Intuition Behind MLE

Let say we have a coin with bernoulli distribution for heads = ?

We can toss the coin 100 times. Let’s say that it landed on heads 68 times.

What do you think the probability for heads is?

0.68!

You just used the MLE principal

MLE - Maximum Likelihood Estimation

We can also prove that MLE converges to the true parameters when the observation number goes to infinity

Think about the coin example.
If you toss it n=3 times and get 2 heads: pMLE = 0.66

If you make n=1000 and get 705 heads: pMLE = 0.705

Y-DATA
SCHOOL OF DATA SCIENCE

Let’s Go Back to Bayesian Law

Probability of *Dog* given that *x=6.5*

Probability of *x=6.5* given that *y* is Dog

Probability of *y* being Dog

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Probability of *x=6.5*

Y-DATA
SCHOOL OF DATA SCIENCE

Let’s Go Back to Bayesian Law

Probability of *y* being “True” given that *x* is “True”

Probability of *x* being “True” given that *y* is “True”

Probability of *y* being “True”

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Probability of *x* being “True”

Y-DATA
SCHOOL OF DATA SCIENCE

Let's Go Back to Bayesian Law

Posterior

Likelihood

Prior

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Evidence

Y-DATA
SCHOOL OF DATA SCIENCE

MLE - Maximum Likelihood Estimation

If I know the outcome, what is the probability of a certain observation?

Ex: for Dogs and Cats: P(Dogs) = # dogs / (# dogs + # cats)
P(Weight|Dog) = We make a gaussian assumption and Mu and Sigma are calculated based on all observed dogs in the world

Y-DATA
SCHOOL OF DATA SCIENCE

Max A-Posterior (MAP) classifier

$$\hat{y}_{MAP} = argmax_y P(y|x)$$

$$= argmax_y \frac{P(x|y)P(y)}{P(x)}$$

Positive const
Not depend on y

$$= argmax_y P(x|y)P(y)$$

likelihood

Prior

Formulate Dog vs Cat Problem with MAP

Y - category {Dog, Cat}

x - weight

Which is larger:

$P(y=Dog|x=6.5)$ or $P(y=Cat|x=6.5)$?

Max A-Posterior (MAP) classifier

$$\hat{y}_{MAP} = argmax_y P(y|x)$$

$$= argmax_y \frac{P(x|y)P(y)}{P(x)}$$

$$= argmax_y P(x|y)P(y)$$

What is the most probable category y {Cat, Dog} given weight x?

Probability of weight x given category y {Cat, Dog}

Probability of category y {Cat, Dog}

Y-DATA
SCHOOL OF DATA SCIENCE

Formulate Dog vs Cat Problem with MAP

We’ll transform the problem with Bayesian Rule to observed data:

Which is larger:

$p(x=6.5|y=Dog)*p(y=Dog)$ or $p(x=6.5|y=Cat)*p(y=Cat)$?

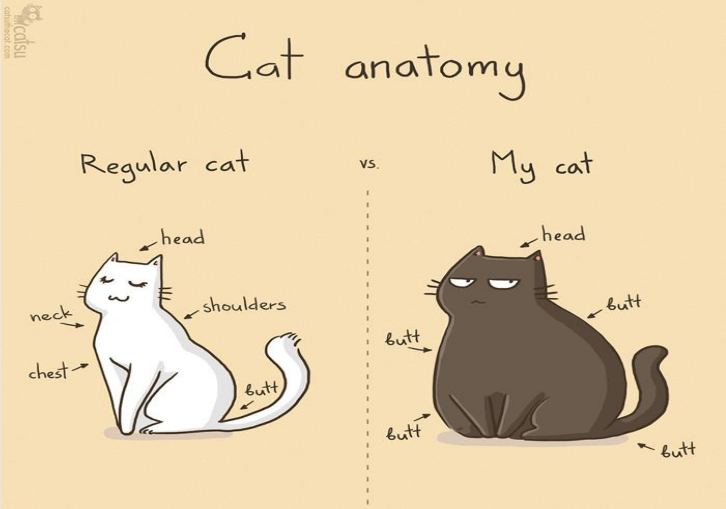
Reminder: We Know That

Dogs - 471M (56%)

Cats - 373M (44%)

p(weight|Cat) ~ N(5,1)

p(weight|Dog) ~ N(10,4)



Gaussian Naive Bayes with MLE

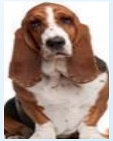
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

$$\Pr(y) = \frac{N_y}{N}$$

p(x=6.5|y=dog) = 0.68
p(y=dog) = 0.56

p(x=6.5|y=cat) = 0.13
p(y=cat) = 0.44

0.68*0.56 > 0.13*0.44 → **It's a dog!**



MLE is a specific case of MAP

In the special case when prior follows a uniform distribution, MAP can be written as:

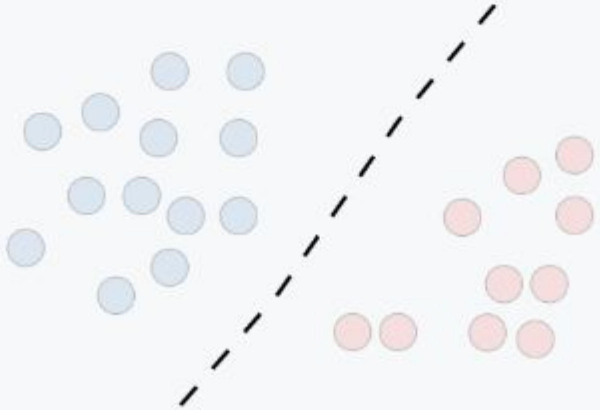
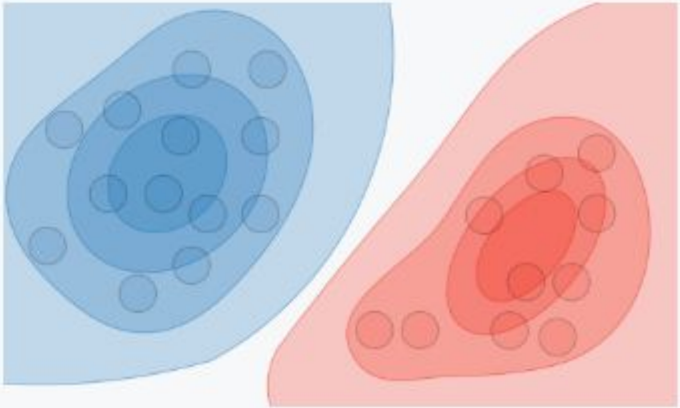
$$\hat{y}_{MAP} = argmax_y P(y|x)$$
$$= argmax_y \frac{P(x|y)P(y)}{P(x)}$$
$$= argmax_y P(x|y)P(y)$$
$$= argmax_y P(x|y).$$

Types of Naive Bayes Classifiers

- ***Gaussian Naive Bayes*** - Used when we are dealing with continuous data and uses Gaussian distribution.
- ***Bernoulli Naive Bayes*** - Used for discrete data, where features are only in binary form.
- ***Multinomial Naive Bayes*** - Widely used classifier for document classification which keeps the count of frequent words present in the documents.
- ***Complement Naive Bayes*** - Used for imbalanced data

Y-DATA
SCHOOL OF DATA SCIENCE

Generative vs Discriminative Models

	Discriminative model	**Generative model**
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Practical Example

Y-DATA
SCHOOL OF DATA SCIENCE

Boy or Girl?

Pregnant woman at week 20. Task: Boy or Girl ?

Measurements:

1. Weight gain (Kg)
2. Avg amount of chocolate eaten weekly (gra

Collect training data

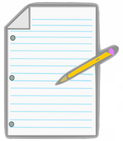
Weight Gain	Chocolate Craving	Gender
10	4	Boy
7	5	Boy
13	1	Boy
11	1	Boy
10	6	Boy
9	0	Boy
6	5	Boy
9	3	Boy
15	5	Boy
1	1	Boy
4	0	Girl
5	1	Girl
10	3	Girl
7	0	Girl
3	1	Girl
7	1.5	Girl
5	1	Girl
3	0	Girl
5	0	Girl
3	0.5	Girl

Y-DATA
SCHOOL OF DATA SCIENCE

Data Processing: Discretization

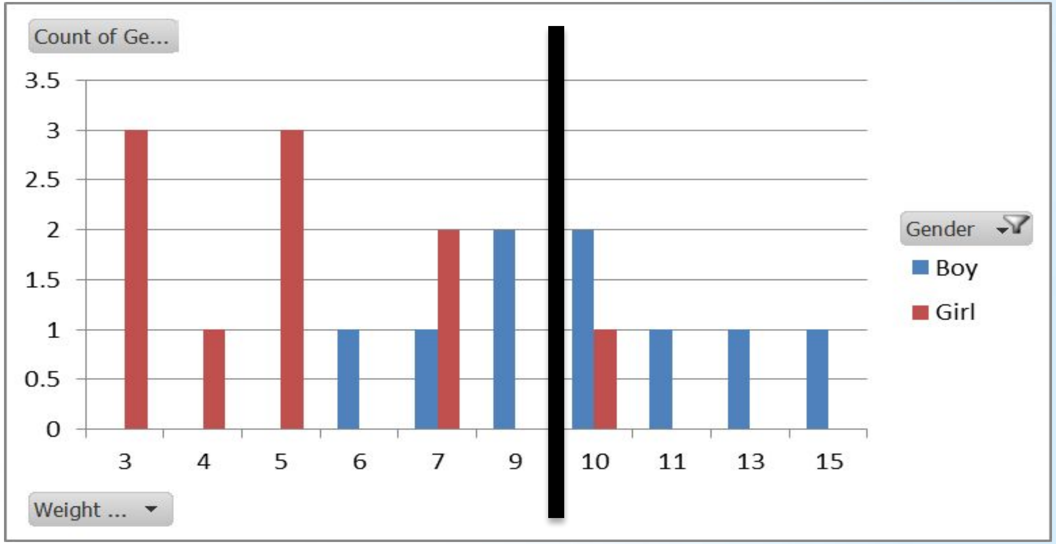
Pregnant woman at week 20.

Task: Boy or Girl ? : Gender (G)



Weight gain: Weight Gain (W)

1. Low if less than 9Kg
2. High if more than 9kg



Y-DATA
SCHOOL OF DATA SCIENCE

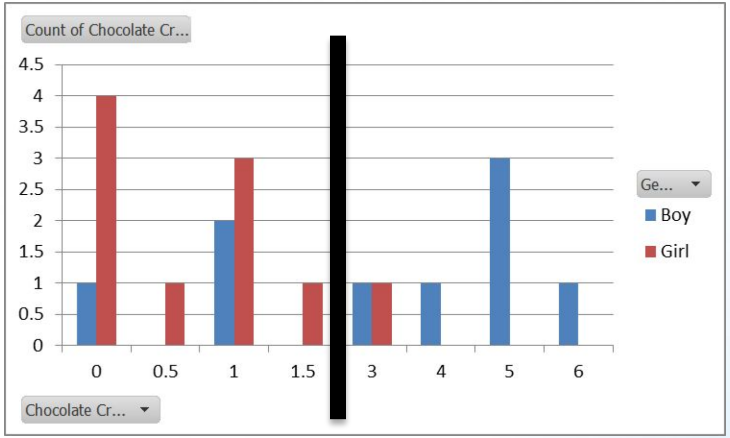
Data Processing: Discretization

Pregnant woman at week 20

Task: Boy or Girl ? : Gender (G)

Avg amount of chocolate eaten weekly (bars): Chocolate Craving (CC)

1. Low if less than 2 bars
2. High if more than 2 bars



We Have Observed Labeled Data

W = Weight Gain	CC = Chocolate Craving	G = Gender
1	1	Boy
0	1	Boy
1	0	Boy
1	0	Boy
1	1	Boy
1	0	Boy
0	1	Boy
1	1	Boy
1	1	Boy
0	0	Boy
0	0	Girl
0	0	Girl
1	1	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl

Y-DATA
SCHOOL OF DATA SCIENCE

Let’s Predict The Outcome

Pregnant Woman:

Weight Gain: 10 kg → 1

Chocolate Craving: 1 bar → 0

Argmax gender p(gender|w=1, cc=0) ?

Mathematical Formulation

$\text{Argmax}_G\ p(G \mid W, CC)$

$= \text{Argmax}_G\ p(W, CC \mid G) * P(G) / P(W, CC)$

$= \text{Argmax}_G\ p(W, CC \mid G) * P(G)$

*** Assuming independence**

$= \text{Argmax}_G\ p(W \mid G) * P(CC \mid G) * P(G)$

Likelihood

	W = 0	W = 1
G = Boy	0.3	0.7
G = Girl	0.9	0.1

	CC = 0	CC = 1
G = Boy	0.4	0.6
G = Girl	0.9	0.1

G = Boy	0.5
G = Girl	0.5

Weight Gain	Chocolate Craving	Gender
1	1	Boy
0	1	Boy
1	0	Boy
1	0	Boy
1	1	Boy
1	0	Boy
0	1	Boy
1	1	Boy
1	1	Boy
0	0	Boy
0	0	Girl
0	0	Girl
1	1	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl

Y-DATA
SCHOOL OF DATA SCIENCE

Likelihood

	W = 0	W = 1
G = Boy	0.3	0.7
G = Girl	0.9	0.1

	CC = 0	CC = 1
G = Boy	0.4	0.6
G = Girl	0.9	0.1

G = Boy	0.5
G = Girl	0.5

Let's Go Back to Our Woman: w=1, cc=0

Argmax G P(W|G) * P(CC | G) * P(G)

Boy →

P(W=1|G=boy) * P(CC = 0 | G=boy) * P(G=boy)=

0.7*0.4*0.5 = 0.14

Girl →

P(W=1|G=girl) * P(CC = 0 | G=girl) * P(G=girl)=

0.1*0.9*0.5= 0.045

0.14 > 0.045 → **It's a boy!**

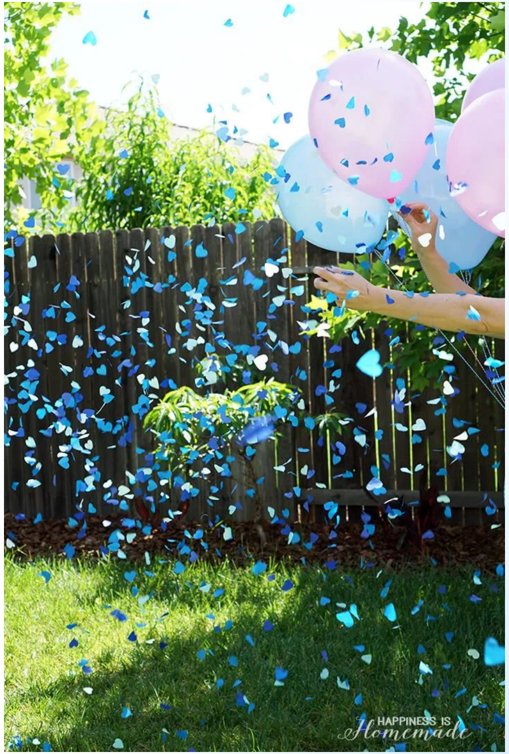
Y-DATA
SCHOOL OF DATA SCIENCE

Likelihood

	W = 0	W = 1
G = Boy	0.3	0.7
G = Girl	0.9	0.1

	CC = 0	CC = 1
G = Boy	0.4	0.6
G = Girl	0.9	0.1

G = Boy	0.5
G = Girl	0.5



Likelihood

	W = 0	W = 1
G = Boy	0.3	0.7
G = Girl	0.9	0.1

	CC = 0	CC = 1
G = Boy	0.4	0.6
G = Girl	0.9	0.1

Let's Go Back to Our Woman: w=1, cc=0

This is not P(boy|W,CC)!!! What do we need to add to calculate it?

Boy →

P(W=1|G=boy) * P(CC = 0 | G=boy) * ~~P(G=boy)=~~

0.7*0.4*0.5 = 0.14

Girl →

P(W=1|G=girl) * P(CC = 0 | G=girl) * ~~P(G=girl)=~~

0.1*0.9*0.5= 0.045

Since Probabilities are equal

0.14 > 0.045 → **It's a boy!**

Are W and CC Independent? Or Could We be Naive?

- Conditional independence:

P(W, CC | G) = P(W|G) * P(CC|G)

W and CC are independent given G

Each feature has its own conditional probability table:

Question: P(W=0|G=Girl) * P(CC=0|G=Girl) =? P(W=0,CC=0|G=Girl)

Y-DATA
SCHOOL OF DATA SCIENCE

Are W and CC Independent?

	W = 0	W = 1
G = Boy	0.3	0.7
G = Girl	0.9	0.1

	CC = 0	CC = 1
G = Boy	0.4	0.6
G = Girl	0.9	0.1

	W=0, CC=0	W=0, CC=1	W=1, CC=0	W=1, CC=1
G= Boy	0.1	0.2	0.3	0.4
G= Girl	0.9	0	0	0.1

Weight Gain	Chocolate Craving	Gender
1	1	Boy
0	1	Boy
1	0	Boy
1	0	Boy
1	1	Boy
1	0	Boy
0	1	Boy
1	1	Boy
1	1	Boy
0	0	Boy
0	0	Girl
0	0	Girl
1	1	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl

P(W=0|G=Girl) * P(CC=0|G=Girl) ? P(W=0,CC=0|G=Girl)

Y-DATA
SCHOOL OF DATA SCIENCE

Are W and CC Independent?

Weight Gain	Chocolate Craving	Gender
1	1	Boy
0	1	Boy
1	0	Boy
1	0	Boy
1	1	Boy
1	0	Boy
0	1	Boy
1	1	Boy
1	1	Boy
0	0	Boy
0	0	Girl
0	0	Girl
1	1	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl
0	0	Girl

	W = 0	W = 1
G = Boy	0.3	0.7
G = Girl	0.9	0.1

	CC = 0	CC = 1
G = Boy	0.4	0.6
G = Girl	0.9	0.1

	W=0, CC=0	W=0, CC=1	W=1, CC=0	W=1, CC=1
G= Boy	0.1	0.2	0.3	0.4
G= Girl	0.9	0	0	0.1

0.9* 0.9 != 0.9 → **Weight and Chocolate Consumption are Dependant**

Let’s Repeat Without Naive Assumption

	W=0, CC=0	W=0, CC=1	W=1, CC=0	W=1, CC=1
G= Boy	0.1	0.2	0.3	0.4
G= Girl	0.9	0	0	0.1

Let’s Go Back to Our Woman: w=1, cc=0

Argmax G P(W, CC | G) * P(G)

Boy →

P(W=1,CC = 0 | G=boy) * P(G=boy)=

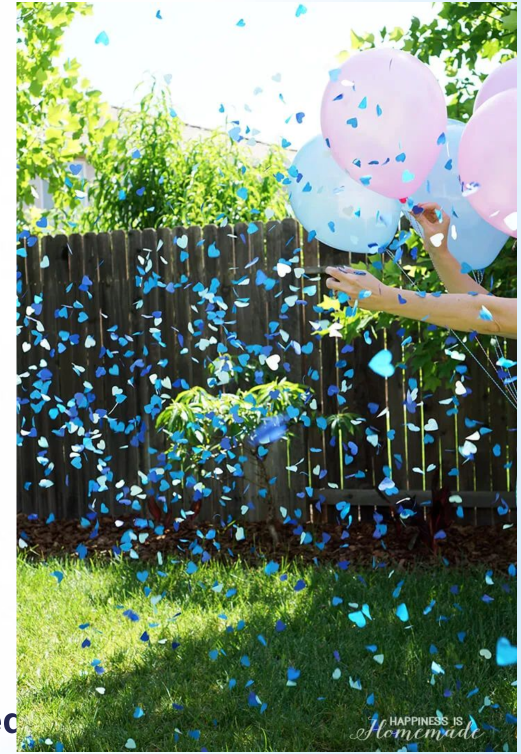
0.3*0.5 = 0.15

Girl →

P(W=1,CC = 0 | G=girl) * P(G=girl)=

0*0.5= 0

0.15 > 0→ **Still a boy (but could have changed**



Y-DATA
SCHOOL OF DATA SCIENCE

Naive Bayes Language Model

Based on: CIS 391 – Introduction to Artificial Intelligence

Example: Classifying Spam Mails

- Desired output ‘Label’ an article to one of three categories (multiclass)
    - Spam
    - Family
    - Work

Y-DATA
SCHOOL OF DATA SCIENCE

Data

- Collection of 1.5M emails (documents) where
  - 500K labeled as spam
  - 500K labeled as family
  - 500K labeled as work

Hi,
How are you?
You must buy the new nike shoes
Here is a link:
http://www.nike.com/bla

Multinomial Naive Bayes for Text Classification

- $P(w_i|c)$ is the conditional probability of word w_i occurring in document of class *c*
- $P(c)$ is the prior probability of a document occurring in class *c*
- n_d is the number of such tokens in *d*
- $P(c|d)$ is the probability of class given document

The model: Argmax$_c$ P(c|d) = Argmax$_c$ p(d|c) * p(c) | p(d) = Argmax$_c$ p(c)*p(d|c)

p(d|c) = p(w1,w2,w3,w4….wn|c)
= p(w1|c)*p(w1|w2,c)*p(w3|w1,w2,c)....p(wn|w1,...wn-1, c)
Assuming independence = p(w1|c)*p(w2|c)*p(w3|c)...(wn|c)

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i | c_j)$$

Multinomial Naive Bayes for Text Classification

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

Instead of multiplication of probabilities, use **sum of logs** to avoid underflow

Y-DATA
SCHOOL OF DATA SCIENCE

How will we calculate the parameters?

- **From training corpus, extract** ***Vocabulary***

- **Calculate required $P(c_j)$ and $P(w_k \mid c_j)$ terms**
  - For each c_j in C do :

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

In our case?

How will we calculate the parameters?

- **From training corpus, extract** ***Vocabulary***

- **Calculate required $P(c_j)$ and $P(w_k \mid c_j)$ terms**
  - For each c_j in C do :

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

In our case: P(spam) = P(family) = p(work) = 1/3

How will we calculate the parameters?

- For each word w_k in *Vocabulary*

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

What happens if a word didn't appear in any document?

Y-DATA
SCHOOL OF DATA SCIENCE

Discounting - Handling Missing Data

Example: Lidstone discounting

$$p_{Lid}(x) = \frac{C(x) + \lambda}{|S| + \lambda |X|}$$

Vocabulary size

For lambda=1 it’s called laplace discounting

Bonus: We can prove that **pLid(x)> pMLE(x)**

How will we calculate the parameters?

- For each word w_k in $\mathit{Vocabulary}$

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Naive Bayes as a Generative Model

Notice that we can use the model we build to classify if an email is spam or not in order to build a generative model! This model is called "Language Model"

$P(c_j)$

How?

We pick a class by tossing a bernoulli coin with

Let's say we got c=spam

Generative Story

Now we generate the text accordingly.

We pick the first word by sampling p(w|c=spam)

Code

Y-DATA
SCHOOL OF DATA SCIENCE

Let’s Think About This Together

1. What are the main hyper-parameters?
2. Can it work for Multi-class data?
3. Does it handle Categorical data?
4. Does it handle missing data?
5. Is it sensitive to outliers?
6. What if some features are correlated?
7. Is it Interpretable?
8. Can it be parallelized?
9. Speed of training
10. Speed of prediction

Y-DATA
SCHOOL OF DATA SCIENCE

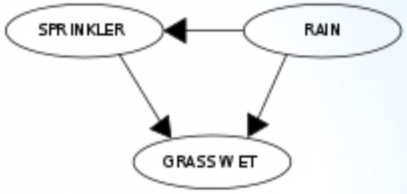
Let’s Think About This Together

1. What are the main hyper-parameters? Distribution type, smoothing parameter
2. Can it work for Multi-class data? Yes
3. Does it handle Categorical data? Yes
4. Does it handle missing data? Yes
5. Is it sensitive to outliers? No
6. What if some features are correlated? Naive assumptions don’t work
7. Is it Interpretable? Yes
8. Can it be parallelized? Yes
9. Speed of training - Fast (Linear Time)
10. Speed of prediction - Fast (Linear Time)

Non-Naive Bayesian Methods

- Bayesian Networks

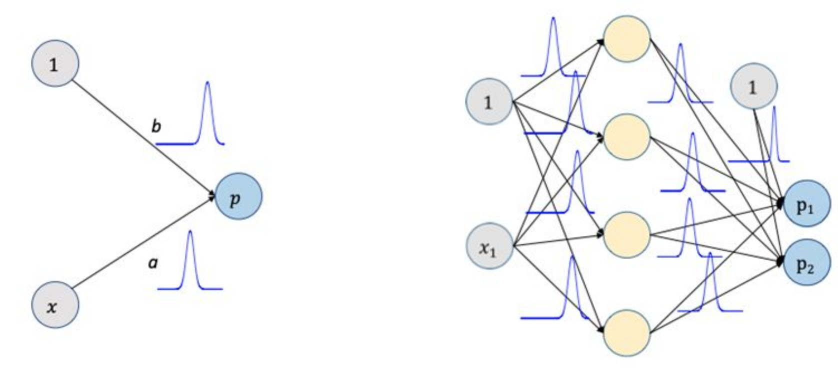
RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99



RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- Deep Bayesian Networks



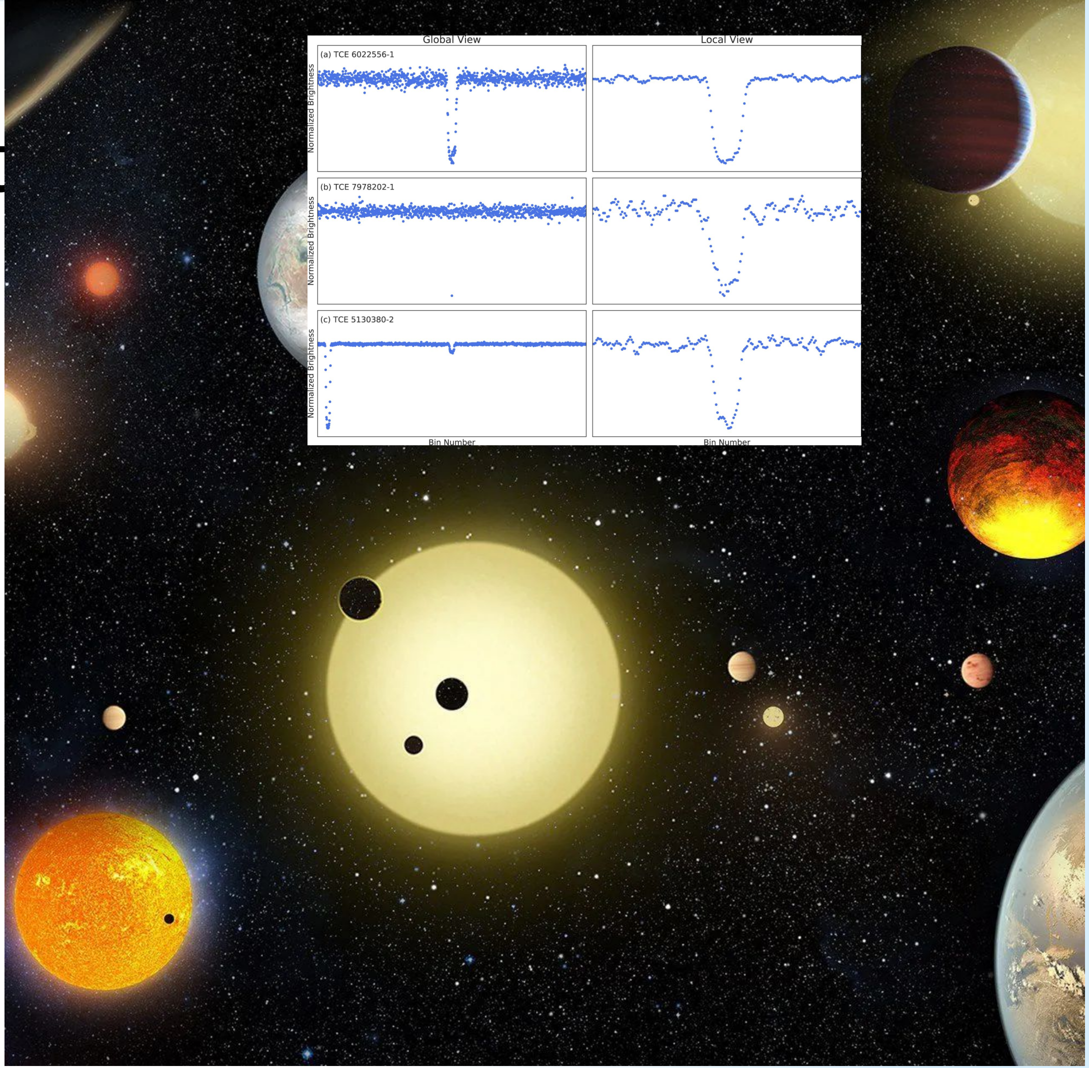
Y-DATA
SCHOOL OF DATA SCIENCE

VESPA Exoplanet

The generative category includes models such as **vespa** (Morton & Johnson 2011; Morton 2012; Morton et al. 2016) that require the class priors $p(y=1)$ and $p(y=0)$ and likelihoods $P(X|y=1)$ and $P(X|y=0)$ to estimate the posterior $P(y=1|X)$ according to the Bayes' theorem:

$$p(y=1|X) = \frac{p(X|y=1)p(y=1)}{\sum_y p(X|y)p(y)} \quad (1)$$

where $y=1$ represents an exoplanet, $y=0$ represents a false positive, and X is a representation of the transit signal. Such generative approaches require the detailed knowledge of the likelihood, $P(X|y)$, and prior, $P(y)$, for each class (exoplanet vs false positive), and class scenario (e.g., BEB). While in general both the likelihood and priors can be learned in a data-driven approach (using ML), **vespa** estimates them by sim-



Summary

Naive Bayes

- Is a generative probabilistic model (the only one we’ll see in SL class)
- Can be used for classification, regression and generation
- Uses MLE and MAP principals

Y-DATA
SCHOOL OF DATA SCIENCE

Pros and Cons

Pros	Cons
1. Very fast 2. Good for big data with big velocity 3. Ignores interactions and therefore needs less data 4. Works well with multiclass problems 5. Works with missing data 6. Can be used as a generative model (e.g. generate text)	1. Small dataset leads to instability (probabilities can be 0 or 1 with high variance) 2. Doesn't perform well for imbalanced datasets 3. Continuous features require binning or assumption of a distribution