10-301/601: Introduction to Machine Learning Lecture 17 - Naïve Bayes

Henry Chai

10/31/22

Front Matter

- Announcements:
  - HW6 released 10/27, due 11/4 at 11:59 PM
    - Only two late days allowed on HW6
    - **HW6 recitation on Wednesday 11/2; next lecture is on Friday, 11/4**
  - Exam 2 on 11/10
    - All topics between Lecture 8 and Lecture 17 (today's lecture) are in-scope
    - Exam 1 content may be referenced but will not be the primary focus of any question
  - Fill out the mid-semester survey, due 11/2
    - As of 9 AM this morning, only 228/405 ≈ 56%

Q & A:

Who were you going to come dressed as?

Recall: Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is
$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$
- The partial derivative of the log-likelihood is
$$\frac{N_1}{\hat{\phi}} - \frac{N_0}{1-\hat{\phi}} = 0 \rightarrow \frac{N_1}{\hat{\phi}} = \frac{N_0}{1-\hat{\phi}}$$
$$\rightarrow N_1\left(1-\hat{\phi}\right) = N_0\hat{\phi} \rightarrow N_1 = \hat{\phi}(N_0 + N_1)$$
$$\rightarrow \hat{\phi} = \frac{N_1}{N_0 + N_1}$$
- where N_1 is the number of 1's in $\left\{x^{(1)}, \dots, x^{(N)}\right\}$ and N_0 is the number of 0's

Poll Question 1:

After flipping your coin 5 times, what is the MLE of your coin?

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1 - \phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$

- The partial derivative of the log-likelihood is

$$\frac{N_1}{\hat{\phi}} - \frac{N_0}{1 - \hat{\phi}} = 0 \rightarrow \frac{N_1}{\hat{\phi}} = \frac{N_0}{1 - \hat{\phi}}$$

$$\rightarrow N_1\left(1 - \hat{\phi}\right) = N_0 \hat{\phi} \rightarrow N_1 = \hat{\phi}(N_0 + N_1)$$

$$\rightarrow \hat{\phi} = \frac{N_1}{N_0 + N_1}$$

- where N_1 is the number of 1's in $\left\{x^{(1)}, \dots, x^{(N)}\right\}$ and N_0 is the number of 0's

Maximum a Posteriori (MAP) Estimation

- Insight: sometimes we have *prior* information we want to incorporate into parameter estimation
- Idea: use Bayes rule to reason about the *posterior* distribution over the parameters

MLE finds $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ P(D \mid \theta)$

MAP finds $\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\ P(\theta \mid D)$

$$= \underset{\theta}{\operatorname{argmax}}\ \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \underbrace{P(D \mid \theta)}_{\text{likelihood}}\ \underbrace{P(\theta)}_{\text{prior}}$$

$$= \underset{\theta}{\operatorname{argmax}}\ \log P(D \mid \theta) + \log P(\theta)$$

Maximum a Posteriori (MAP) Estimation

1. Specify the *generative story*, i.e., the data generating distribution, including a *prior distribution*

2. Maximize the log-posterior of $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$

$$\ell_{MAP}(\theta) = \log p(\theta) + \sum_{i=1}^{N} \log p\left(x^{(i)} | \theta\right)$$

3. Solve in *closed form*: take partial derivatives, set to 0 and solve

Coin Flipping MAP

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

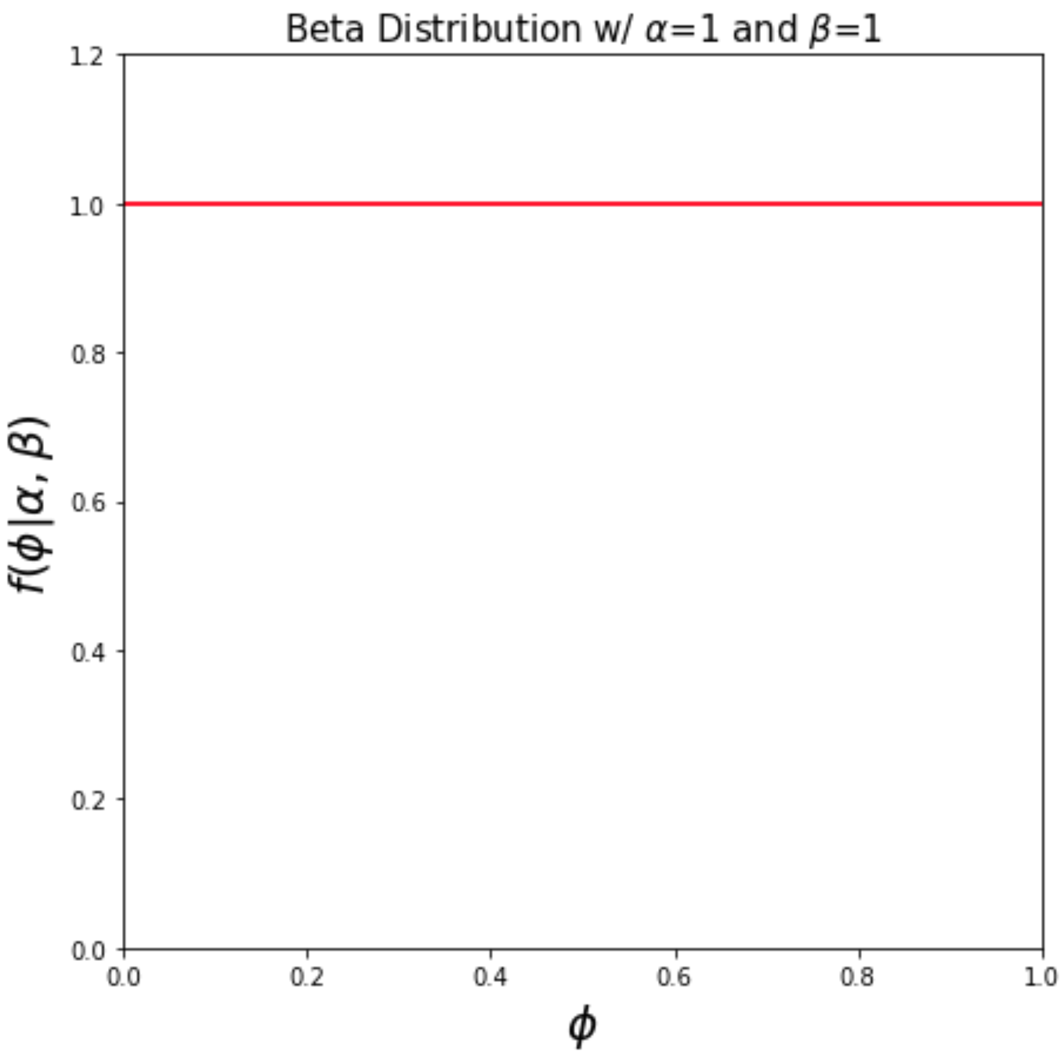
$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

- Assume a Beta prior over the parameter ϕ, which has pdf

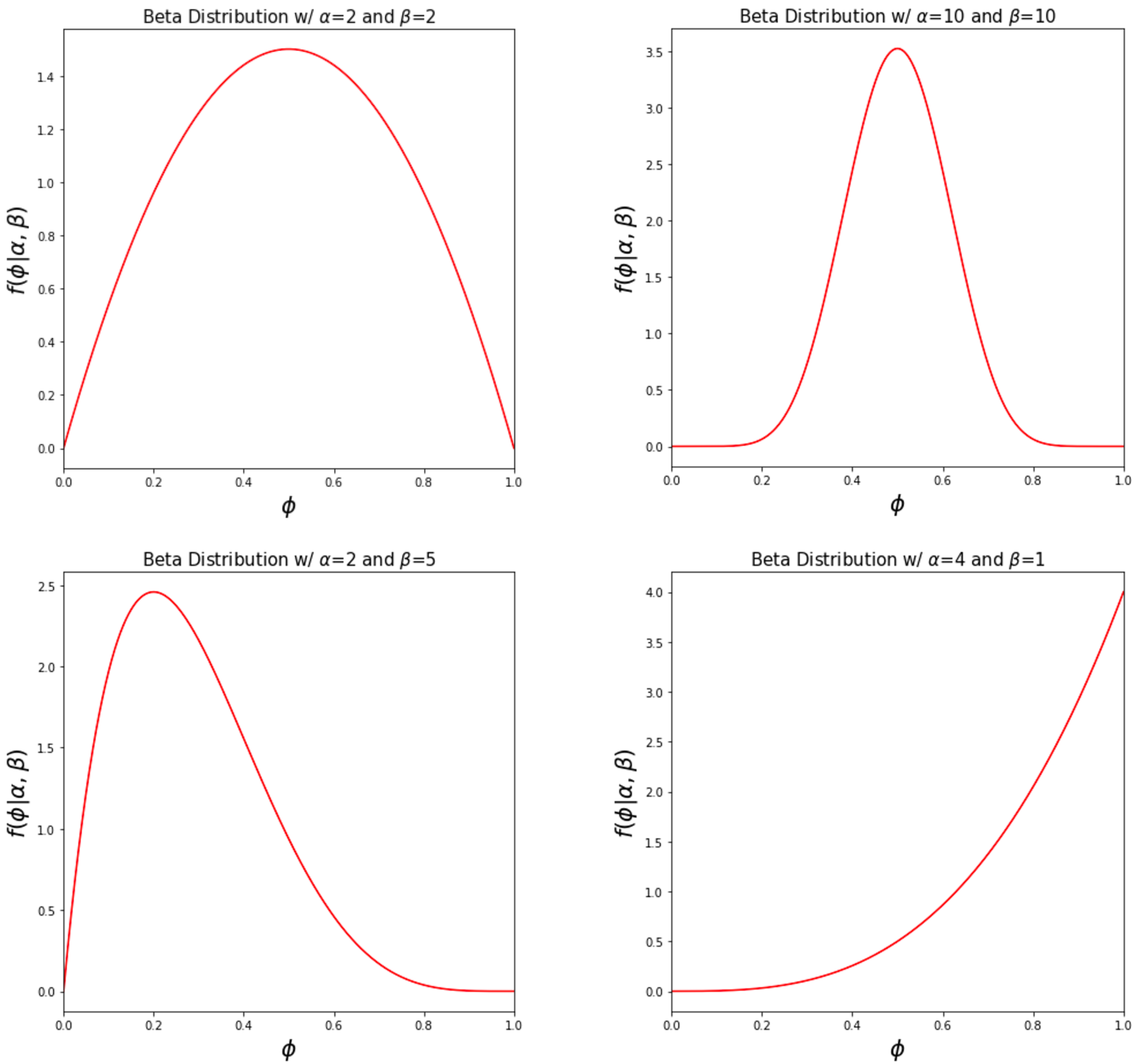
$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \int_0^1 \phi^{\alpha-1}(1-\phi)^{\beta-1} d\phi$ is a normalizing constant to ensure the distribution integrates to 1

Beta Distribution



Beta Distribution



Why use this strange looking Beta prior?

The Beta distribution is the *conjugate prior* for the Bernoulli distribution!

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

- Assume a Beta prior over the parameter ϕ, which has pdf

$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \int_0^1 \phi^{\alpha-1}(1-\phi)^{\beta-1}d\phi$ is a normalizing constant to ensure the distribution integrates to 1

Coin Flipping MAP

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the log-posterior is

$$\ell(\phi) = \log(P(\phi)) + \log(P(D|\phi))$$

$$= \log \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha,\beta)} + \sum_{i=1}^{N} \log \phi^{x^{(i)}}(1-\phi)^{1-x^{(i)}}$$

$$= (\alpha-1)\underline{\log\phi} + (\beta-1)\underline{\log(1-\phi)} - \log B(\alpha,\beta)$$

$$+ \sum_{i=1}^{N}\left(x^{(i)}\log\phi + (1-x^{(i)})\log(1-\phi)\right)$$

$$= (\alpha-1+N_1)\log\phi + (\beta-1+N_0)\log(1-\phi)$$

$$-\log B(\alpha,\beta)$$

where $N_i = \#$ of i's in D

Coin Flipping MAP

- Given N iid samples $\{x^{(1)}, \ldots, x^{(N)}\}$, the partial derivative of the log-posterior is

$$\frac{\partial \ell}{\partial \phi} = \frac{(\alpha - 1 + N_1)}{\phi} - \frac{(\beta - 1 + N_0)}{1 - \phi}$$

$$\vdots$$

$$\rightarrow \hat{\phi}_{MAP} = \frac{(\alpha - 1 + N_1)}{(\beta - 1 + N_0) + (\alpha - 1 + N_1)}$$

- $\alpha - 1$ is a "pseudocount" of the number of 1's (or heads) you've "observed"
- $\beta - 1$ is a "pseudocount" of the number of 0's (or tails) you've "observed"

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 2$ and $\beta = 5$, then

$$\phi_{MAP} = \frac{(2-1+10)}{(2-1+10)+(5-1+2)} = \frac{11}{17} < \frac{10}{12}$$

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 101$ and $\beta = 101$, then

$$\phi_{MAP} = \frac{(101-1+10)}{(101-1+10)+(101-1+2)} = \frac{110}{212} \approx \frac{1}{2}$$

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10 + 2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 1$ and $\beta = 1$, then

$$\phi_{MAP} = \frac{(1 - 1 + 10)}{(1 - 1 + 10) + (1 - 1 + 2)} = \frac{10}{12} = \phi_{MLE}$$

MLE/MAP Learning Objectives

You should be able to…

- Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence
- State the principle of maximum likelihood estimation and explain what it tries to accomplish
- State the principle of maximum a posteriori estimation and explain why we use it
- Derive the MLE or MAP parameters of a simple model in closed form

Text Data

- https://www.nytimes.com/2022/10/13/movies/halloween-ends-review.html
- https://www.nytimes.com/2022/10/20/business/the-spirit-of-halloween.html
- https://www.theonion.com/biden-issues-urgent-warning-for-americans-to-decide-wha-1849597566

Text Data

The TOOTH Book
Dr. Seuss
POCKET ROCKS
SHEREE FITCH
HOP ON POP
By Dr. Seuss
The SIMPLEST SEUSS for YOUNGEST USE
Touch, Feel
THE CAT IN THE HAT
By Dr. Seuss
Go, Dog. Go!
by P.D.Eastman
My Mother?
by P.D. Eastman
Lois Ehlert
RED FISH
BLUE FISH
NEVER touch a
dinosaur!
Chicka Chicka Boom Boom
by Bill Martin Jr
and John Archambault
illustrated by Lois Ehlert
Big Dog... Little Dog
by P.D. Eastman
幼儿启蒙大全
YOU ER QI MENG DA QUAN
FRIENDS COLLECTION
Pooh & Eeyore

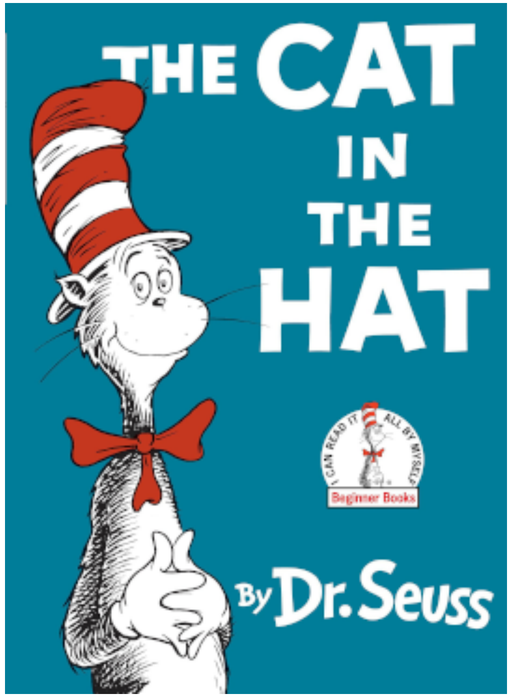
Bag-of-Words Model

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)

Bag-of-Words Model

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)
1	1	0	0	0	0	1

The Cat in the Hat
(by Dr. Seuss)

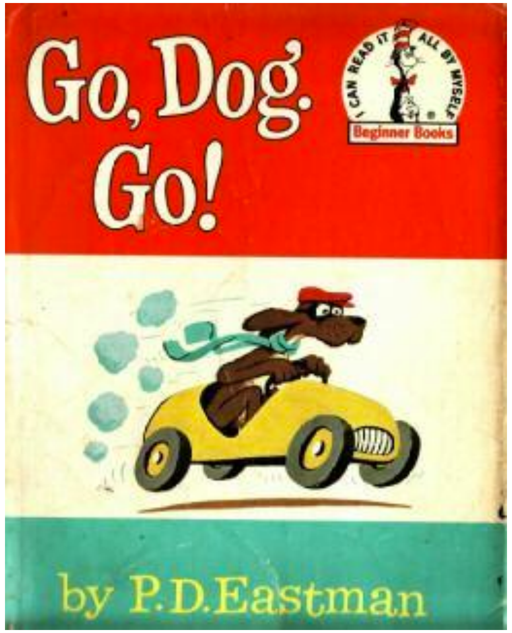


Source: https://en.wikipedia.org/wiki/The_Cat_in_the_Hat#/media/File:The_Cat_in_the_Hat.png

Bag-of-Words Model

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)
1	1	0	0	0	0	1
0	0	1	0	0	0	0

Go, Dog. Go!
(by P. D. Eastman)

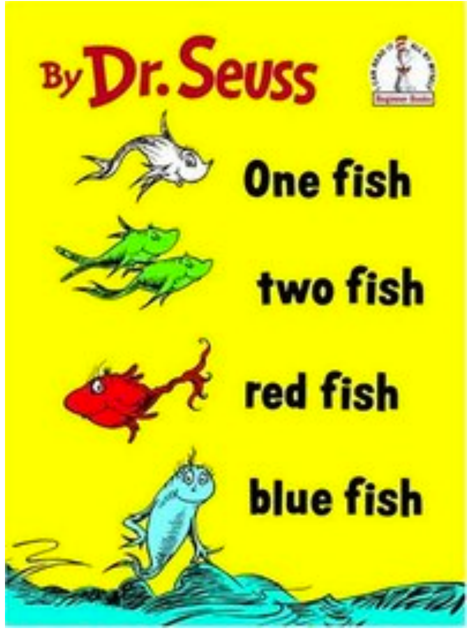


Source: https://en.wikipedia.org/wiki/Go,_Dog._Go!#/media/File:Go_Dog_Go.jpg

Bag-of-Words Model

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)
1	1	0	0	0	0	1
0	0	1	0	0	0	0
0	0	0	1	0	0	1

One Fish, Two Fish, Red Fish, Blue Fish (by Dr. Seuss)

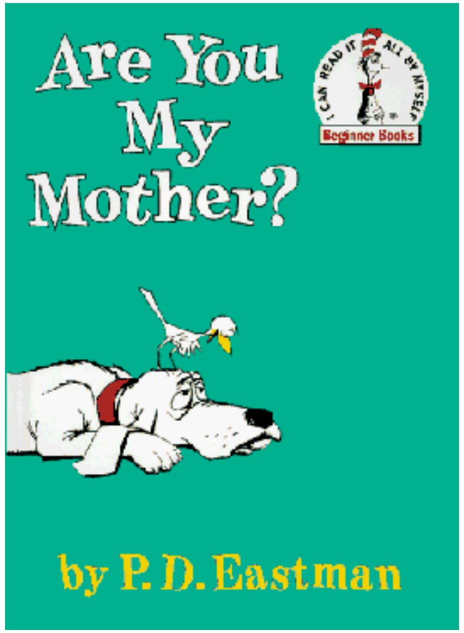


Source: https://en.wikipedia.org/wiki/One_Fish,_Two_Fish,_Red_Fish,_Blue_Fish#/media/File:One_Fish_Two_Fish_Red_Fish_Blue_Fish_(cover_art).jpg

Bag-of-Words Model

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	y (Dr. Seuss)
1	1	0	0	0	0	1
0	0	1	0	0	0	0
0	0	0	1	0	0	1
0	0	0	0	1	0	0

Are You My Mother?
(by P. D. Eastman)



Source: https://en.wikipedia.org/wiki/Are_You_My_Mother%3F#/media/File:Areyoumymother.gif

Recall: Building a Probabilistic Classifier

- Define a decision rule
  - Given a test data point $\boldsymbol{x}'$, predict its label $\hat{y}$ using the posterior distribution $P(Y = y|X = \boldsymbol{x}')$
  - Common choice: $\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y = y|X = \boldsymbol{x}')$
- Model the posterior distribution
  - Option 1 - Model $P(Y|X)$ directly as some function of X (recall: logistic regression)
  - Option 2 - Use Bayes' rule (today!):

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)} \propto P(X|Y)\,P(Y)$$

How hard is modelling $P(X|Y)$?

- Define a decision rule
  - Given a test data point $\boldsymbol{x}'$, predict its label $\hat{y}$ using the posterior distribution $P(Y = y|X = \boldsymbol{x}')$
  - Common choice: $\hat{y} = \underset{y}{\operatorname{argmax}}\, P(Y = y|X = \boldsymbol{x}')$
- Model the posterior distribution
  - Option 1 - Model $P(Y|X)$ directly as some function of X (recall: logistic regression)
  - Option 2 - Use Bayes' rule (today!):

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)} \propto P(X|Y)\,P(Y)$$

How hard is modelling $P(X|Y)$?

x_1 ("hat")	x_2 ("cat")	x_3 ("dog")	x_4 ("fish")	x_5 ("mom")	x_6 ("dad")	$P(X\|Y = 1)$
0	0	0	0	0	0	θ_1
1	0	0	0	0	0	θ_2
1	1	0	0	0	0	θ_3
1	0	1	0	0	0	θ_4

Naïve Bayes Assumption

- ***Assume*** features are conditionally independent given the label: $P(x_1 \cap x_2 \ldots \cap x_D | Y)$

$$P(X|Y) = \prod_{d=1}^{D} P(X_d|Y)$$

- Pros:
  - Significantly reduce our computational costs
  - Helps to combat overfitting
- Cons:
  - This is a really strong (generally incorrect) assumption

$P(Y|X) \propto P(X|Y)P(Y)$

Recipe for Naïve Bayes

- Define a model and model parameters
  - Make the naïve Bayes assumption
  - Assume independent, identically distributed (iid) data
  - Parameters: $\pi = P(Y = 1)$, $\theta_{d,y} = P(X_d = 1|Y = y)$

Example: 2 labels
6 features
12 parameters

- Write down an objective function
  - Maximize the log-likelihood
- Optimize the objective w.r.t. the model parameters
  - Solve in *closed form*: take partial derivatives, set to 0 and solve

Setting the Parameters via MLE

$$\ell_{\mathcal{D}}(\pi, \boldsymbol{\theta}) = \log P\left(\mathcal{D} = \{\boldsymbol{x}^{(1)}, y^{(1)}, \dots, \boldsymbol{x}^{(N)}, y^{(N)}\} \middle| \pi, \boldsymbol{\theta}\right)$$

$$= \log \prod_{i=1}^{N} P(x^{(i)}, y^{(i)}; \pi, \theta)$$

$$\left(P(A \cap B) = P(A|B)P(B) \right)$$

$$= \log \prod_{i=1}^{N} P(x^{(i)} | y^{(i)}; \theta) P(y^{(i)}; \pi)$$

$$= \log \prod_{i=1}^{N} \left(\prod_{d=1}^{D} P(x_d^{(i)} | y^{(i)}; \theta_{d,1}, \theta_{d,0}) \right) P(y^{(i)}; \pi)$$

$$= \sum_{i=1}^{N} \left(\sum_{d=1}^{D} \log P(x_d^{(i)} | y^{(i)}; \theta_{d,1}, \theta_{d,0}) \right) P(y^{(i)}; \pi)$$

Setting the Parameters via MLE

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {}^{N_{Y=1}}/_{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Binary features
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$
  - $\hat{\theta}_{d,y} = {}^{N_{Y=y,\, X_d=1}}/_{N_{Y=y}}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$

Poll Question 2: Given this dataset, what is the MLE of π?

Poll Question 3: Given this dataset, what is the MLE of $\theta_{3,1}$?

x_1	x_2	x_3	y
1	0	1	0
0	1	0	1
0	1	1	1
0	0	1	0
1	0	1	0
1	0	1	1

A. 0/6
B. 1/6
C. 2/6
D. 3/6
E. 4/6 = 2/3
F. 5/6
G. 6/6
H. 7/6 **(TOXIC)**

Bernoulli Naïve Bayes

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {}^{N_{Y=1}}/_{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Binary features
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$
  - $\hat{\theta}_{d,y} = {}^{N_{Y=y,\, X_d=1}}/_{N_{Y=y}}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$

Multinomial Naïve Bayes

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {N_{Y=1}}/{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Discrete features (X_d can take on one of K possible values)
  - $X_d | Y = y \sim \text{Categorical}\left(\theta_{d,1,y}, \dots, \theta_{d,K-1,y}\right)$
  - $\hat{\theta}_{d,k,y} = {N_{Y=y,\, X_d=k}}/{N_{Y=y}}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=k}$ = # of data points with label y and feature $X_d = k$

Gaussian Naïve Bayes

- Binary label
  - $Y \sim \text{Bernoulli}(\pi)$
  - $\hat{\pi} = {}^{N_{Y=1}}/_{N}$
    - N = # of data points
    - $N_{Y=1}$ = # of data points with label 1
- Real-valued features
  - $X_d | Y = y \sim \text{Gaussian}\left(\mu_{d,y}, \sigma^2_{d,y}\right)$
  - $\hat{\mu}_{d,y} = \frac{1}{N_{Y=y}} \underline{\sum_{n: y^{(n)}=y} x_d^{(n)}}$
  - $\hat{\sigma}^2_{d,y} = \frac{1}{N_{Y=y}} \underline{\sum_{n: y^{(n)}=y} \left(x_d^{(n)} - \hat{\mu}_{d,y}\right)^2}$
    - $N_{Y=y}$ = # of data points with label y

Multiclass Gaussian Naïve Bayes

- Discrete label (Y can take on one of M possible values)
  - $Y \sim \text{Categorical}(\pi_1, \dots, \pi_M)$
  - $\hat{\pi}_m = {N_{Y=m}}/{N}$
    - N = # of data points
    - $N_{Y=m}$ = # of data points with label m
- Real-valued features
  - $X_d | Y = y \sim \text{Gaussian}\left(\mu_{d,y}, \sigma^2_{d,y}\right)$
  - $\hat{\mu}_{d,y} = \frac{1}{N_{Y=y}} \sum_{n: y^{(n)}=y} x_d^{(n)}$
  - $\hat{\sigma}^2_{d,y} = \frac{1}{N_{Y=y}} \sum_{n: y^{(n)}=y} \left(x_d^{(n)} - \hat{\mu}_{d,y}\right)^2$
    - $N_{Y=y}$ = # of data points with label y

Visualizing Gaussian Naïve Bayes

- Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width
0	4.3	3.0
0	4.9	3.6
0	5.3	3.7
1	4.9	2.4
1	5.7	2.8
1	6.3	3.3
1	6.7	3.0

Deleted two of the four features, so that input space is 2D

Visualizing Gaussian Naïve Bayes (2 classes)

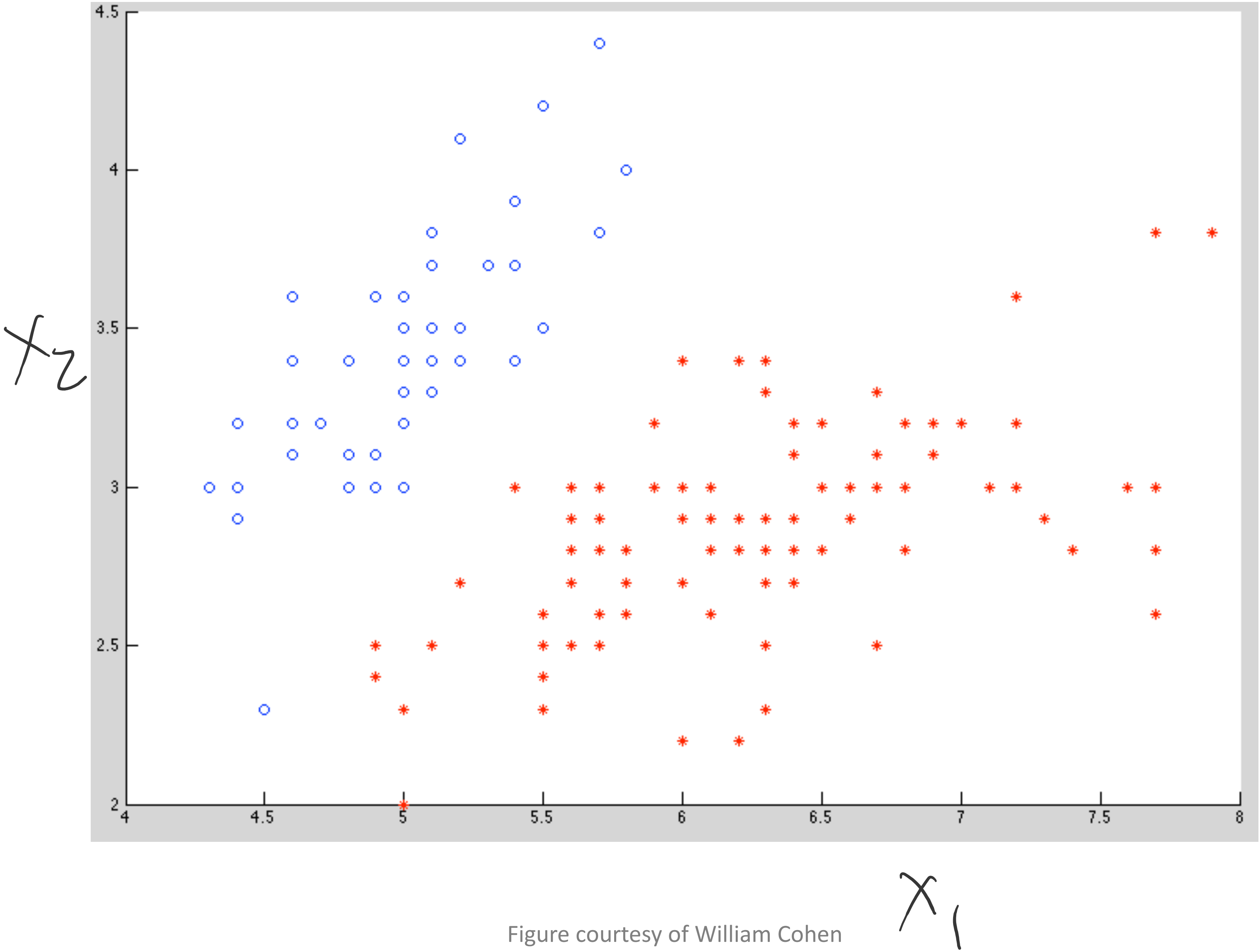


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes)

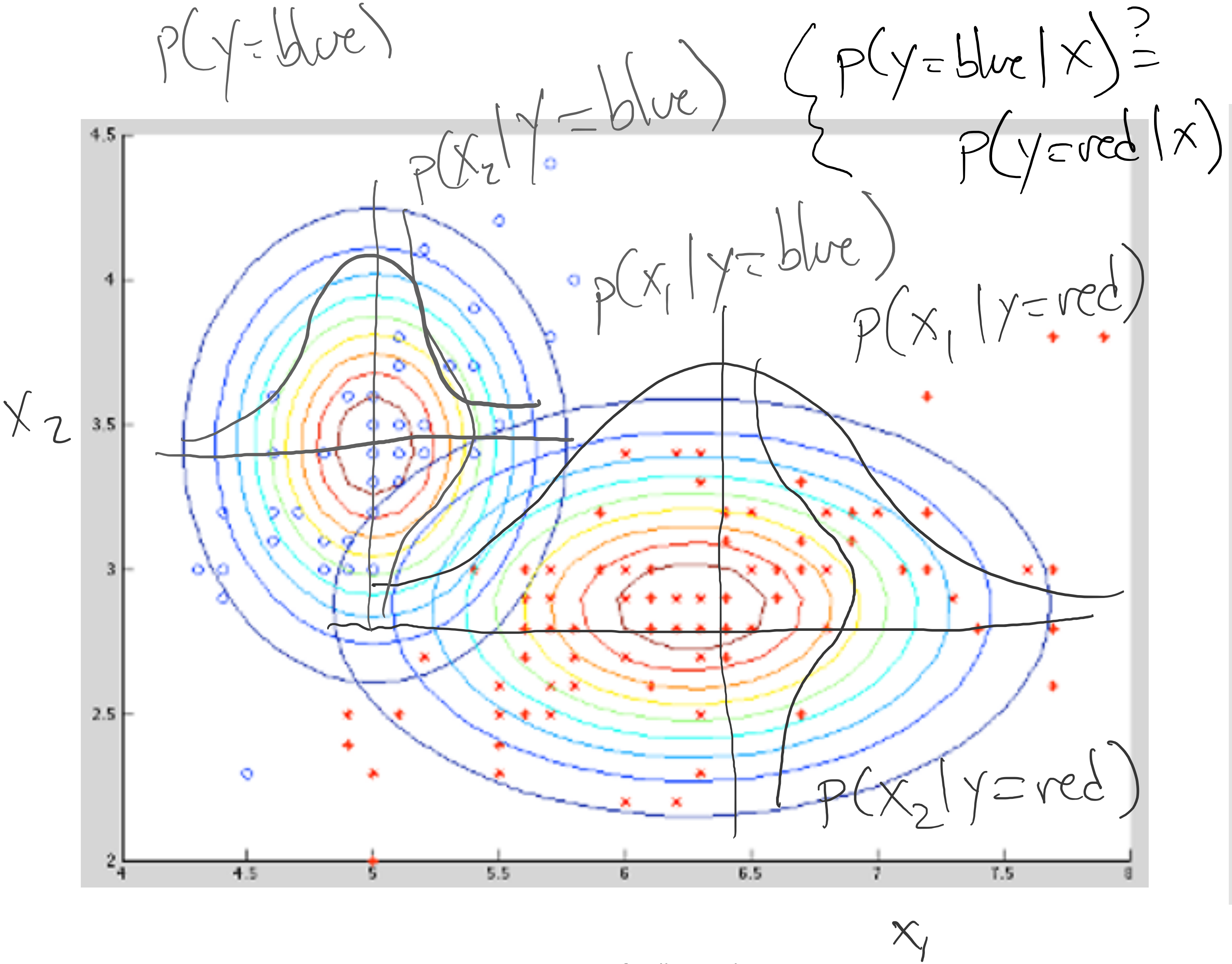


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes, equal variances)

linear

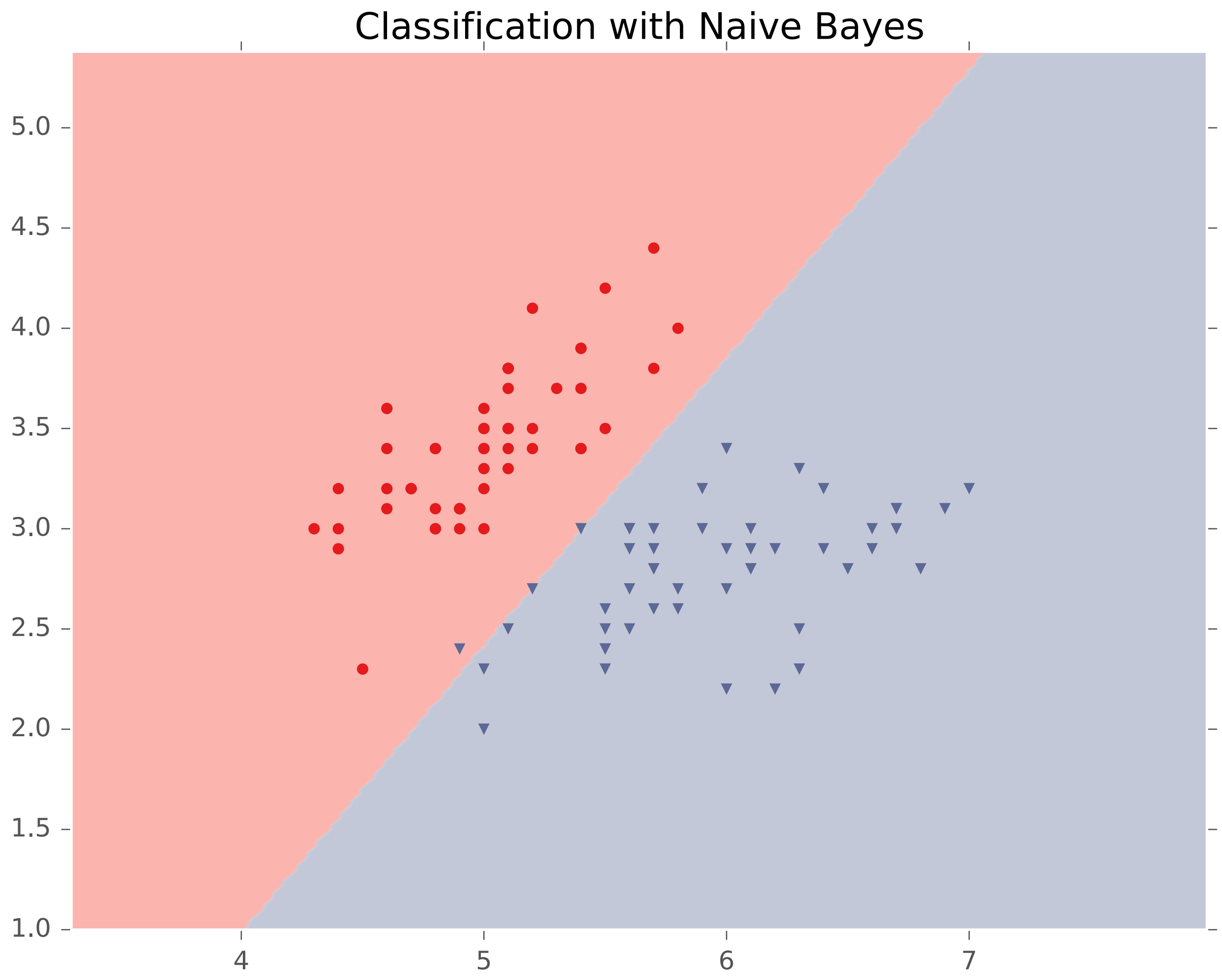


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

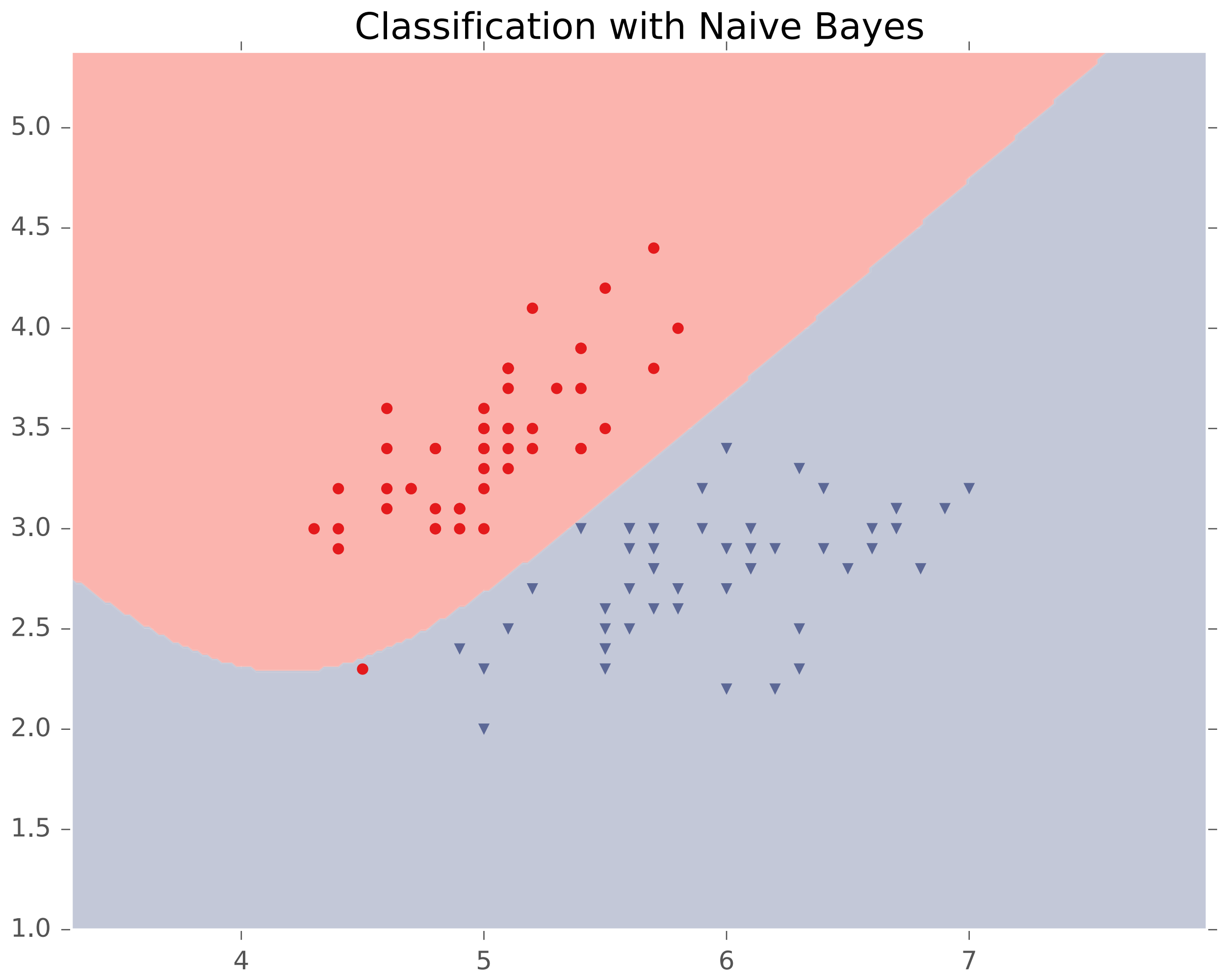


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (3 classes, equal variances)

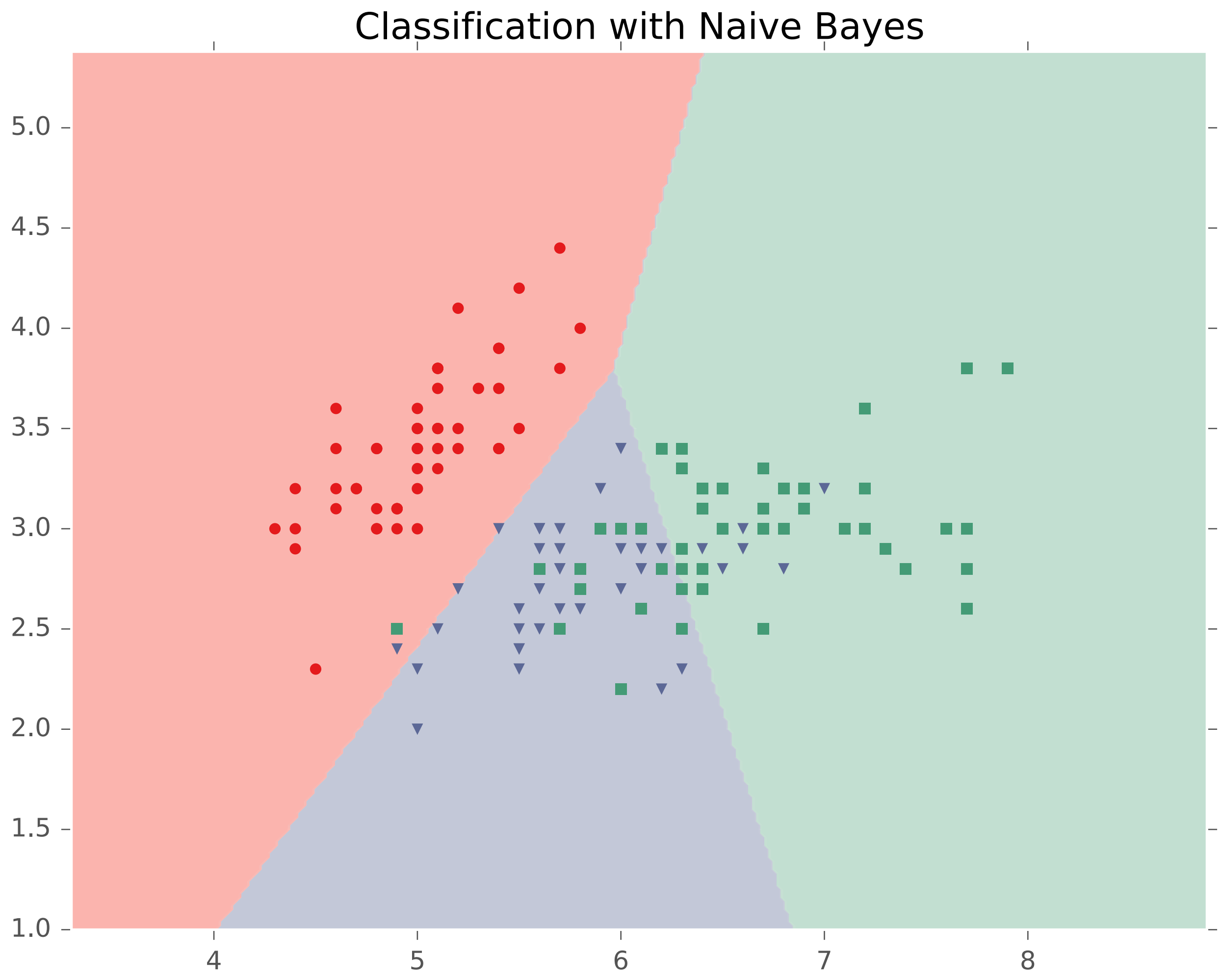


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (3 classes, learned variances)

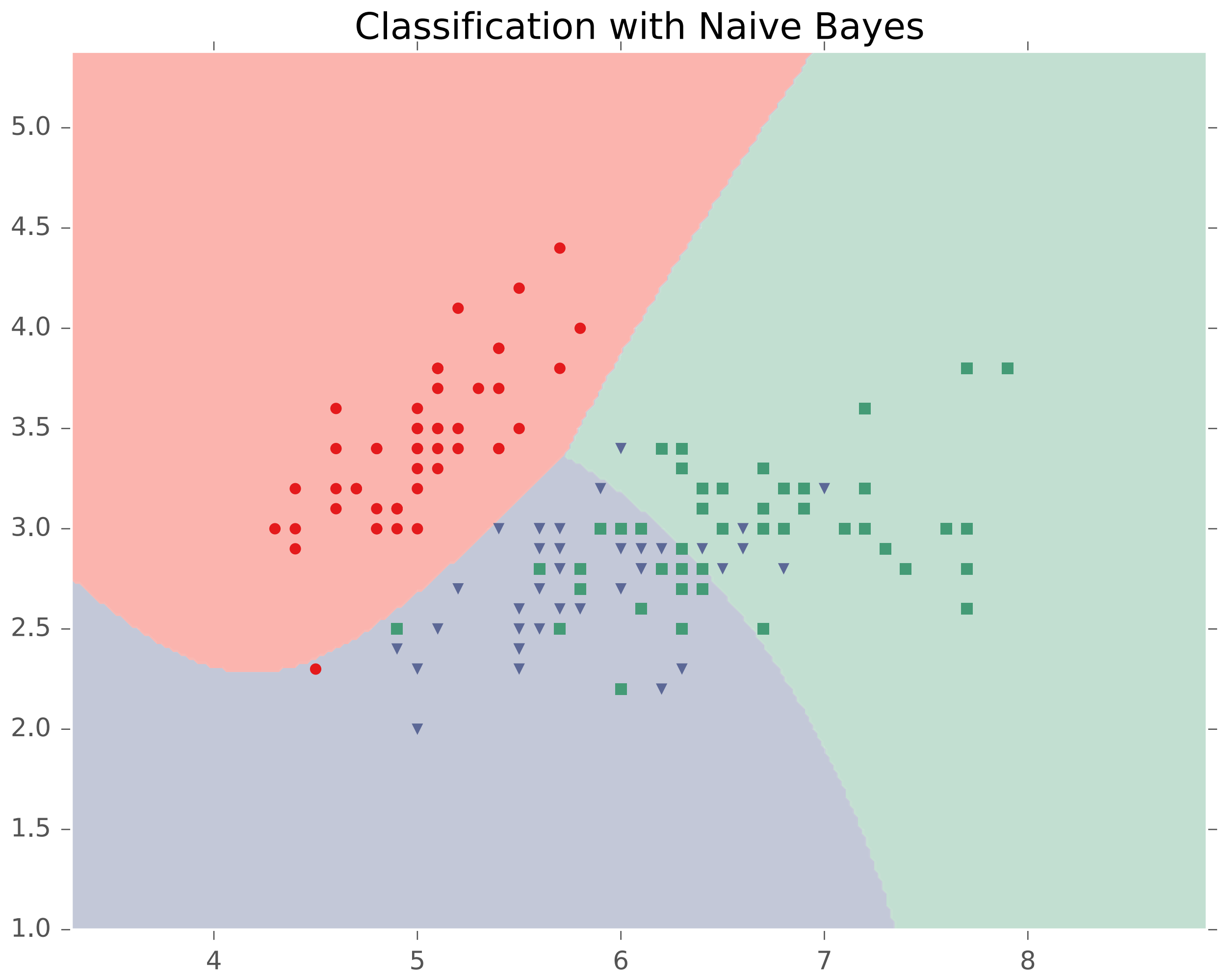


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

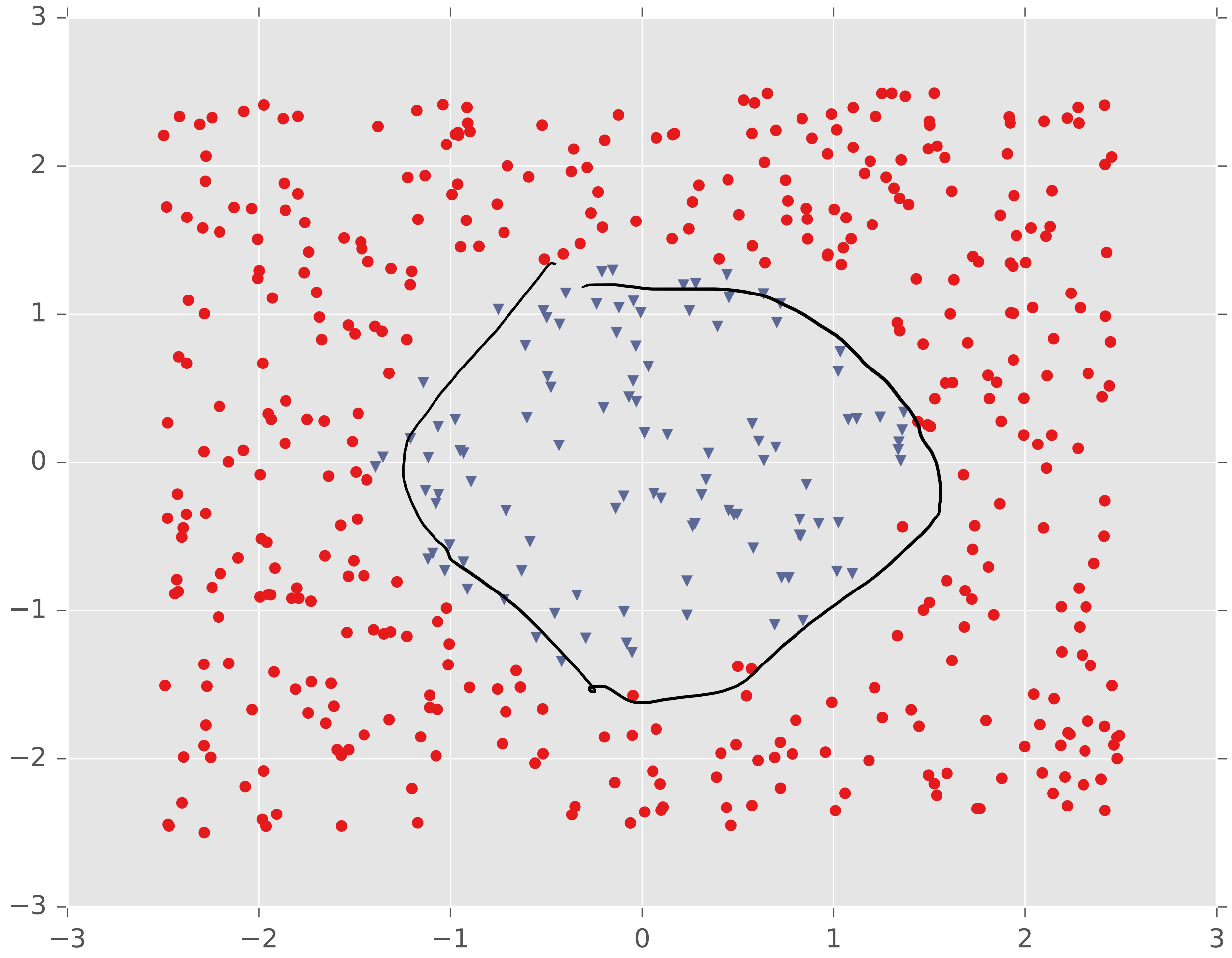


Figure courtesy of Matt Gormley

p(red) >> p(blue)

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

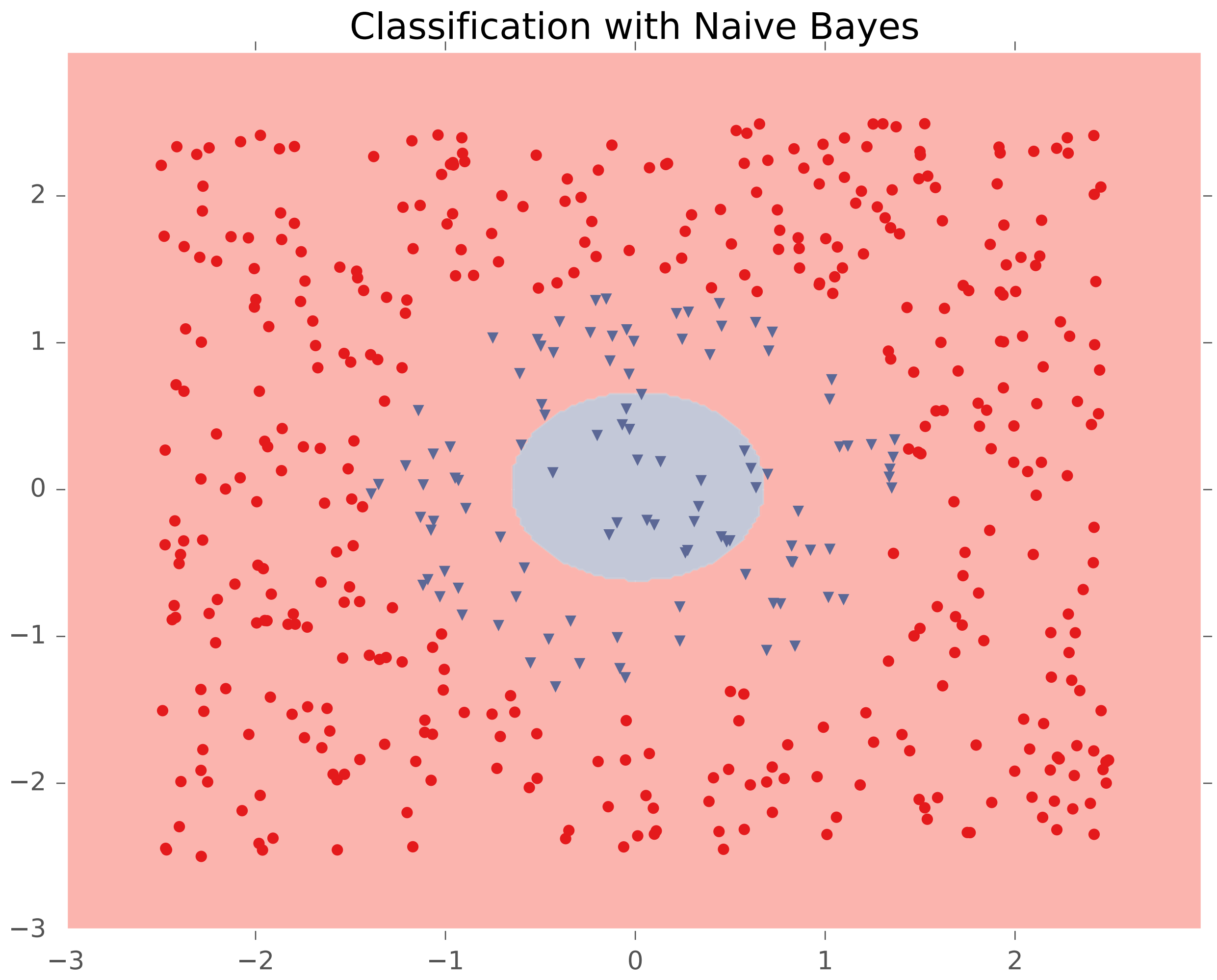


Figure courtesy of Matt Gormley

Naïve Bayes Learning Objectives

You should be able to…

- Write the generative story for Naive Bayes
- Create a new naïve Bayes classifier using your favorite probability distribution as the event model
- Apply the principle of maximum likelihood estimation (MLE) to learn the parameters of Bernoulli naïve Bayes
- Motivate the need for MAP estimation through the deficiencies of MLE
- Apply the principle of maximum a posteriori (MAP) estimation to learn the parameters of Bernoulli naïve Bayes
- Select a suitable prior for a model parameter
- Describe the tradeoffs of generative vs. discriminative models
- Implement Bernoulli naïve Bayes
- Describe how the variance affects whether a Gaussian naïve Bayes model will have a linear or nonlinear decision boundary