10-701: Introduction to Machine Learning Lecture 6 - Naïve Bayes

Henry Chai

2/5/24

Front Matter

- Announcements:
  - Nothing!
- Recommended Readings:
  - Murphy, Section 3.5

Recall: Coin Flipping

- 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

Example: Beta-Binomial Conjugacy

$$f(\phi|x,\alpha,\beta) = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{p(x|\alpha,\beta)}$$

$$p(x|\alpha,\beta) = \int p(x|\phi)f(\phi|\alpha,\beta)d\phi$$

$$= \int \phi^{x}(1-\phi)^{1-x}\frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}d\phi$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)}\int \phi^{\alpha+x-1}(1-\phi)^{\beta-x}d\phi = \frac{\mathrm{B}(\alpha+x,\beta-x+1)}{\mathrm{B}(\alpha,\beta)}$$

Example: Beta-Binomial Conjugacy

$$f(\phi|x,\alpha,\beta) = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{p(x|\alpha,\beta)} = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{\left(\frac{\mathrm{B}(\alpha+x,\beta-x+1)}{\mathrm{B}(\alpha,\beta)}\right)}$$

$$f(\phi|x,\alpha,\beta) = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{\left(\frac{\mathrm{B}(\alpha+x,\beta-x+1)}{\mathrm{B}(\alpha,\beta)}\right)}$$

$$= \frac{\phi^{x}(1-\phi)^{1-x}\frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}}{\left(\frac{\mathrm{B}(\alpha+x,\beta-x+1)}{\mathrm{B}(\alpha,\beta)}\right)}$$

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

$$= f(\phi|\alpha+x,\beta+(1-x))$$

Beta-Binomial MAP

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

$$\ell(\phi) = \log f\big(\phi \big| \alpha + x^{(1)} + x^{(2)} + \cdots x^{(N)},$$
$$\big(\beta + (1 - x^{(1)}) + (1 - x^{(2)}) + \cdots + (1 - x^{(N)})\big)\big)$$

$$= \log f(\phi | \alpha + N_1, \beta + N_0)$$

where N_i is the number of i's observed in the samples

$$= \log \frac{\phi^{\alpha + N_1 - 1}(1 - \phi)^{\beta + N_0 - 1}}{\mathrm{B}(\alpha, \beta)}$$
$$= (\alpha + N_1 - 1) \log \phi + (\beta + N_0 - 1) \log(1 - \phi) - \log \mathrm{B}(\alpha, \beta)$$

Beta-Binomial MAP

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

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

$$\vdots$$

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

- $\alpha - 1$ is a "pseudocount" of the number of 1's you've "observed"
- $\beta - 1$ is a "pseudocount" of the number of 0's 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 ($N_0 = 2$):

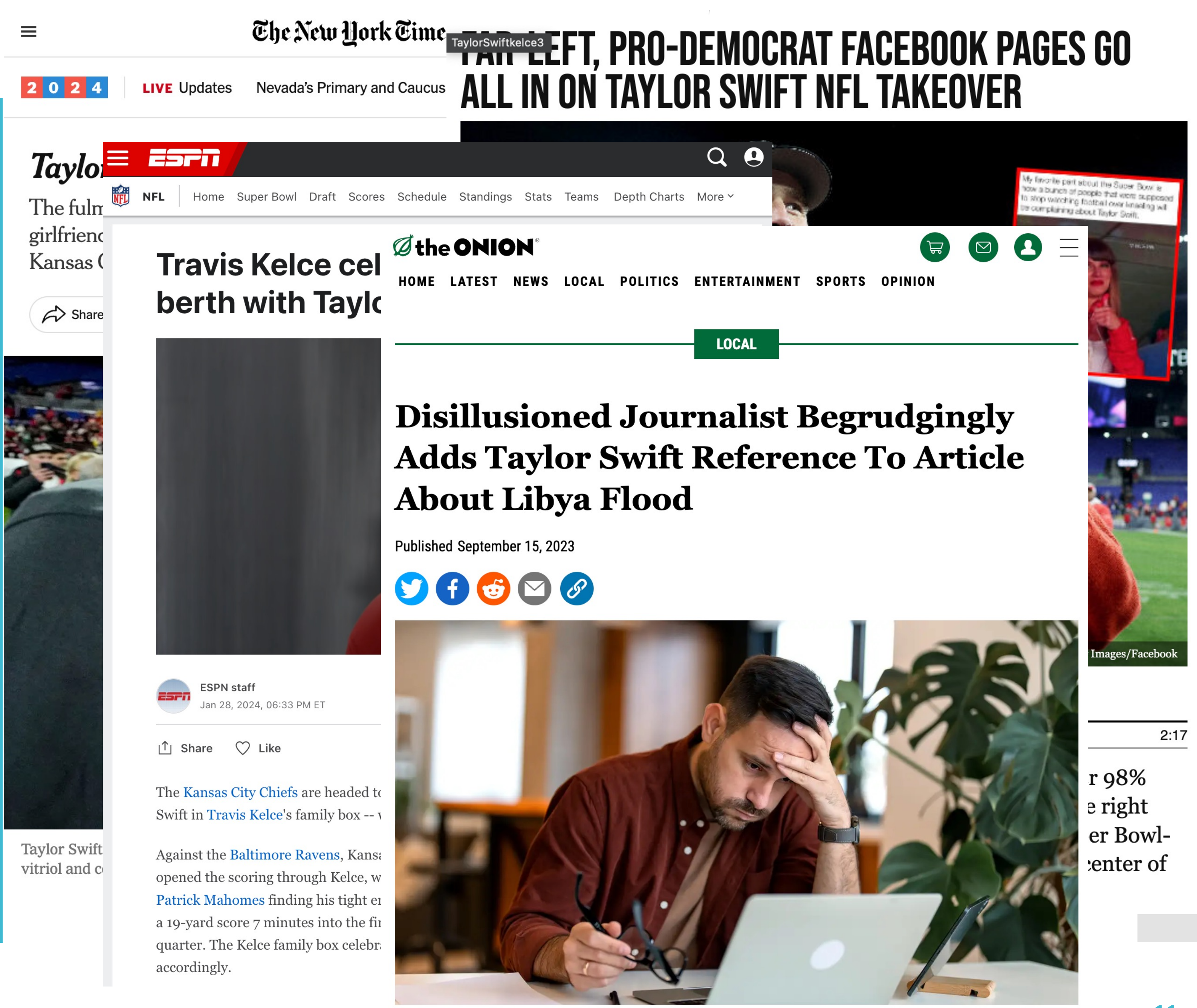
$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

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

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

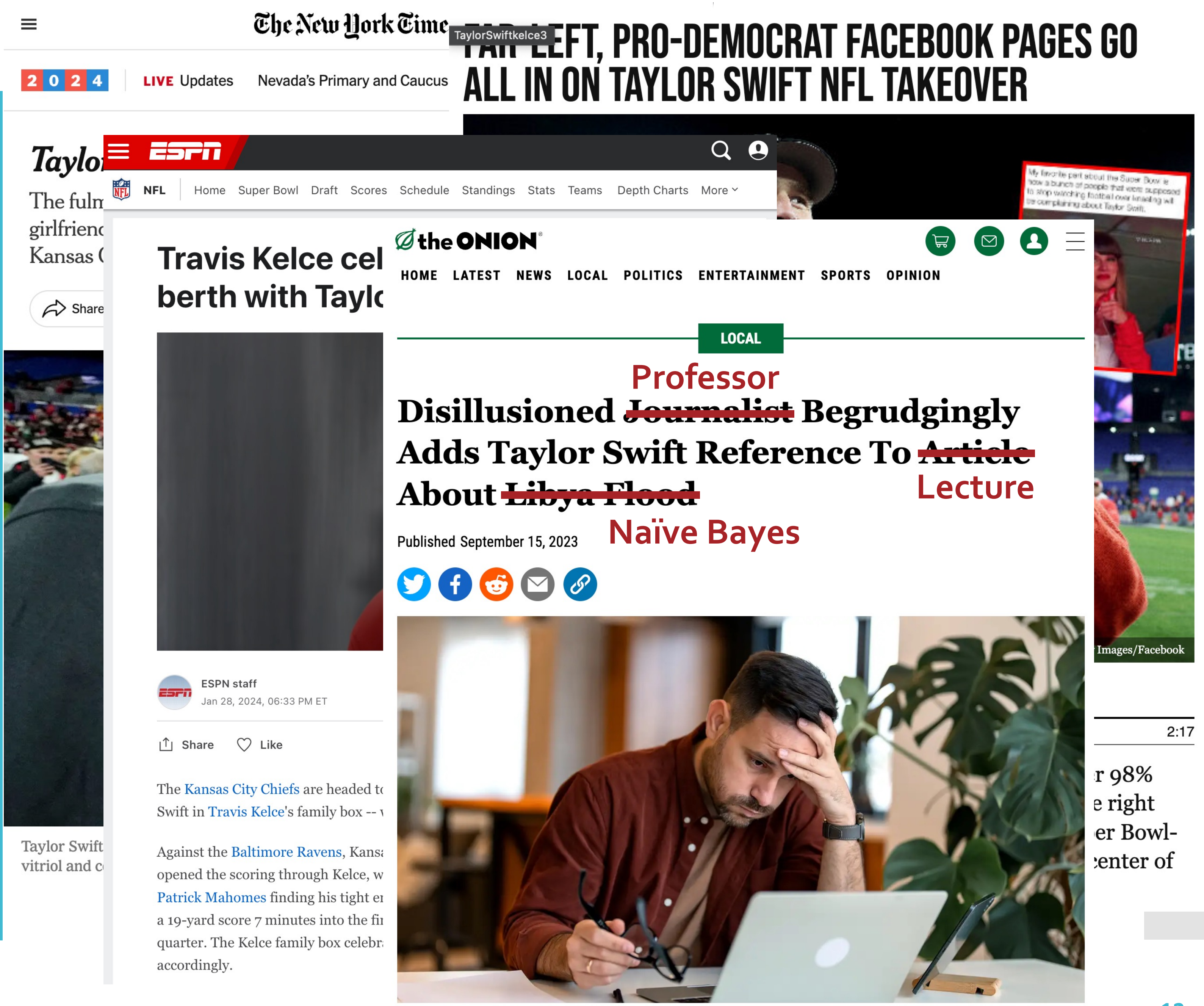
Text Data

- https://www.nytimes.com/2024/01/30/us/politics/taylor-swift-travis-kelce-trump.html
- https://www.breitbart.com/entertainment/2024/01/30/far-left-pro-democrat-facebook-pages-go-all-in-on-taylor-swift-nfl-takeover/
- https://www.espn.com/nfl/story/_/id/39395830/travis-kelce-taylor-swift-afc-championship/
- https://www.theonion.com/disillusioned-journalist-begrudgingly-adds-taylor-swift-1850843119



Text Data

- https://www.nytimes.com/2024/01/30/us/politics/taylor-swift-travis-kelce-trump.html
- https://www.breitbart.com/entertainment/2024/01/30/far-left-pro-democrat-facebook-pages-go-all-in-on-taylor-swift-nfl-takeover/
- https://www.espn.com/nfl/story/_/id/39395830/travis-kelce-taylor-swift-afc-championship/
- https://www.theonion.com/disillusioned-journalist-begrudgingly-adds-taylor-swift-1850843119



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
Lois Ehlert
RED FISH
BLUE FISH
by P.D. Eastman
The CRAYONS' Book of FEELINGS
Chicka Chicka Boom Boom
by Bill Martin Jr and John Archambault
illustrated by Lois Ehlert
幼儿启蒙大全
YOU ER QI MENG DA QUAN
Big Dog... Little Dog
by P.D. Eastman
FRIENDS COLLECTION
Pooh & Eeyore

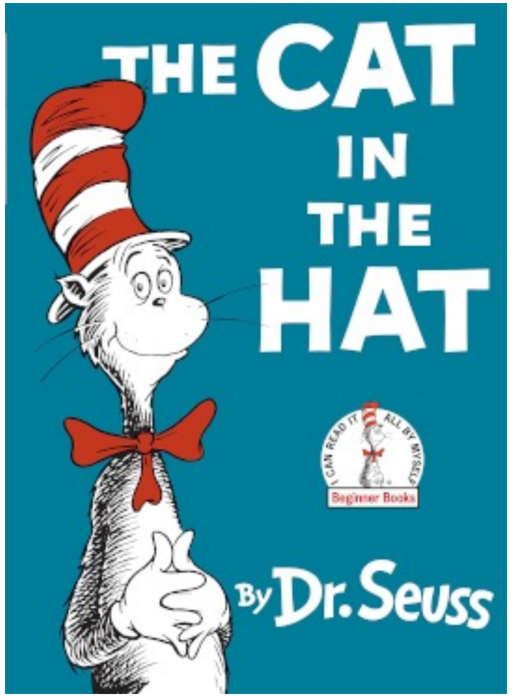
Bag-of-Words Model

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)

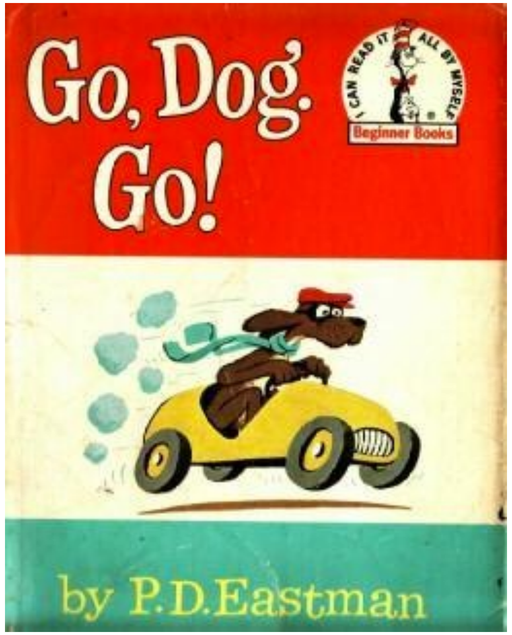


Source: https://en.wikipedia.org/wiki/The_Cat_in_the_Hat#/media/File:The_Cat_in_the_Hat.png

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)

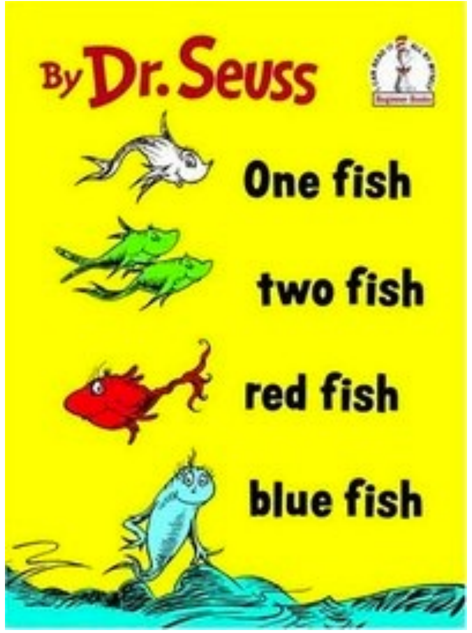


Source: https://en.wikipedia.org/wiki/Go,_Dog._Go!#/media/File:Go_Dog_Go.jpg

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)

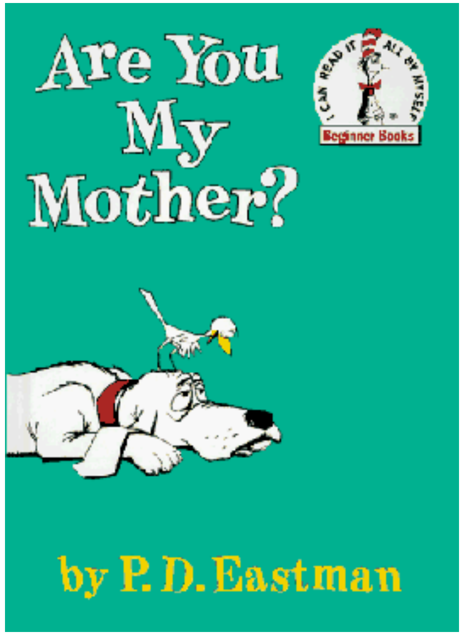


Source: https://en.wikipedia.org/wiki/One_Fish,_Two_Fish,_Red_Fish,_Blue_Fish#/media/File:One_Fish_Two_Fish_Red_Fish_Blue_Fish_(cover_art).jpg

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

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 (Wednesday)
  - 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 (later)
  - 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")	$P(X\|Y = 1)$	$P(X\|Y = 0)$
0	0	0	0	0	0	θ_1	θ_{64}
1	0	0	0	0	0	θ_2	θ_{65}
1	1	0	0	0	0	θ_3	θ_{66}
1	0	1	0	0	0	θ_4	θ_{67}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	1	1	1	1	$1 - \sum_{i=1}^{63} \theta_i$	$1 - \sum_{i=64}^{126} \theta_i$

Naïve Bayes Assumption

- ***Assume*** features are conditionally independent given the label:

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

- Pros:
  - Significantly reduces computational complexity
  - Also reduces model complexity, combats overfitting
- Cons:
  - Is a strong, often illogical assumption
    - We'll see a relaxed version of this later in the semester when we discuss Bayesian networks

General Recipe for Machine Learning

- Define a model and model parameters
- Write down an objective function
- Optimize the objective w.r.t. the model parameters

Recipe for Naïve Bayes

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

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

$$= \log \prod_{n=1}^{N} P\left(\boldsymbol{x}^{(n)}, y^{(n)} \middle| \pi, \boldsymbol{\theta}\right) = \log \prod_{n=1}^{N} P\left(\boldsymbol{x}^{(n)} \middle| y^{(n)}, \boldsymbol{\theta}\right) P\left(y^{(n)} \middle| \pi\right)$$

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

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

$$= \sum_{n:y^{(n)}=1} \left(\sum_{d=1}^{D} \log P\left(x_d^{(n)} \middle| \theta_{d,1}\right) \right)$$

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

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 $X_d = 1$

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 $X_d = 1$

Multiclass Bernoulli Naïve Bayes

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

Recall: Fisher Iris Dataset

- 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

Visualizing Gaussian Naïve Bayes (2 classes)

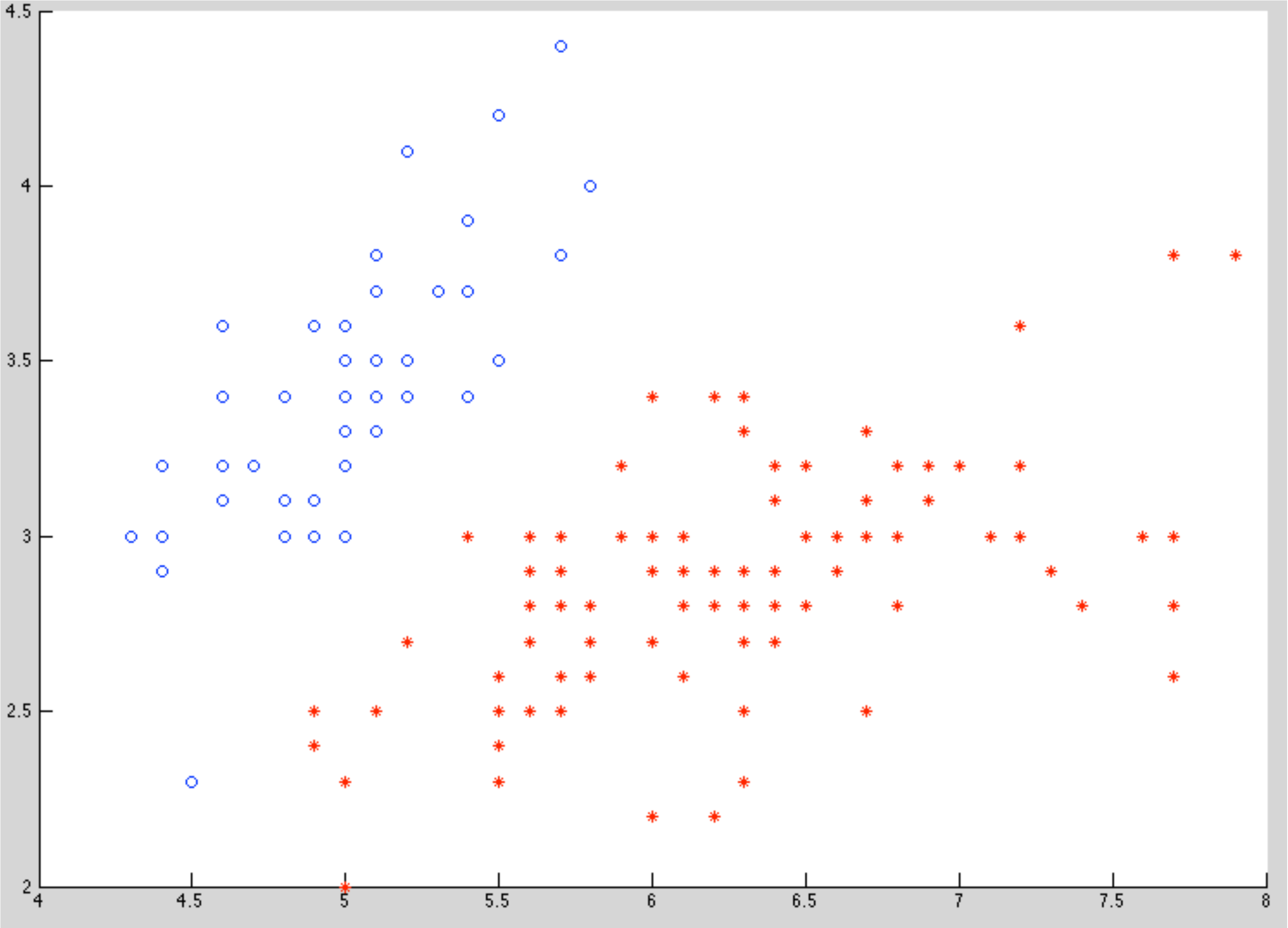


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes)

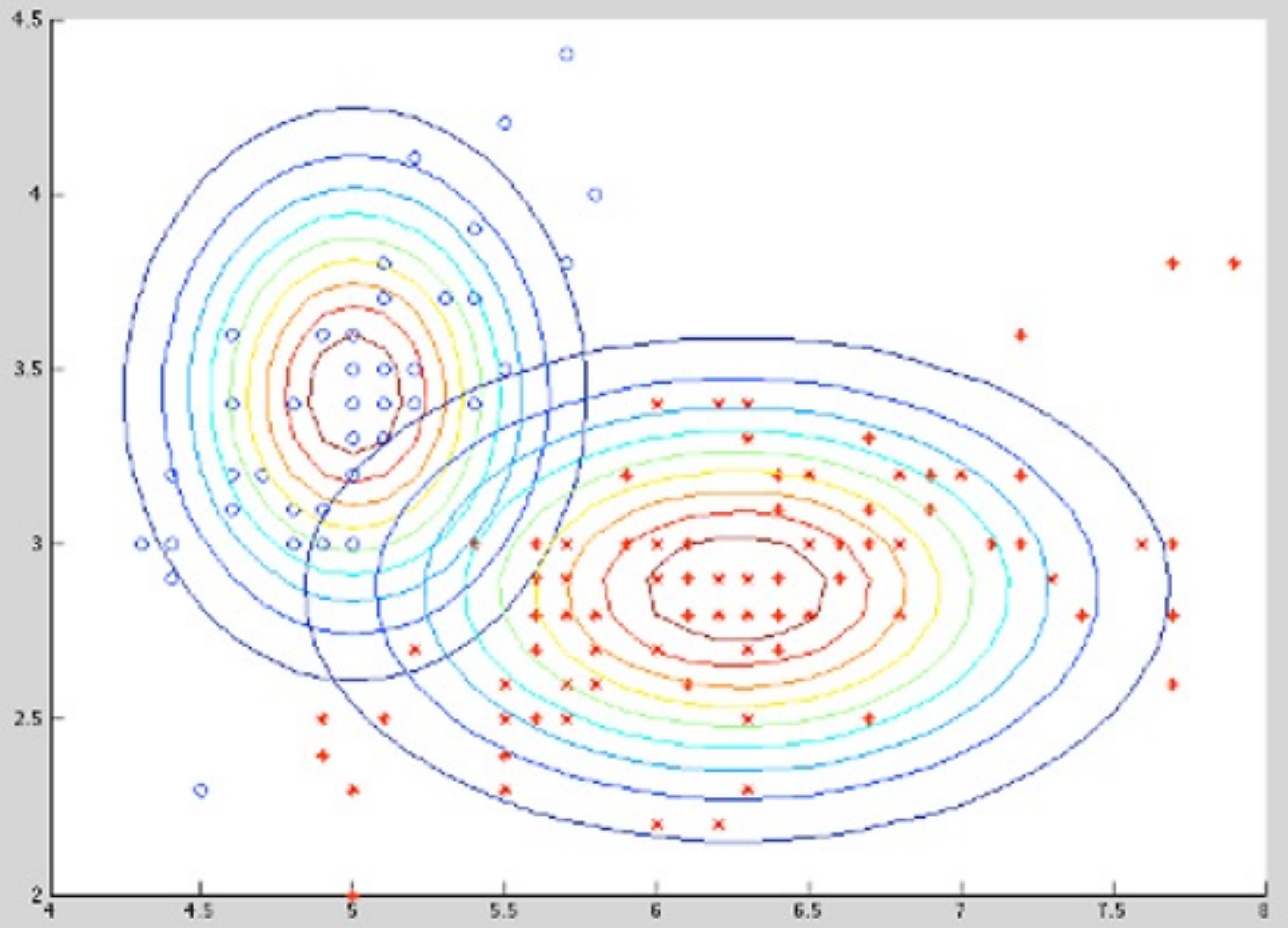


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes, equal variances)

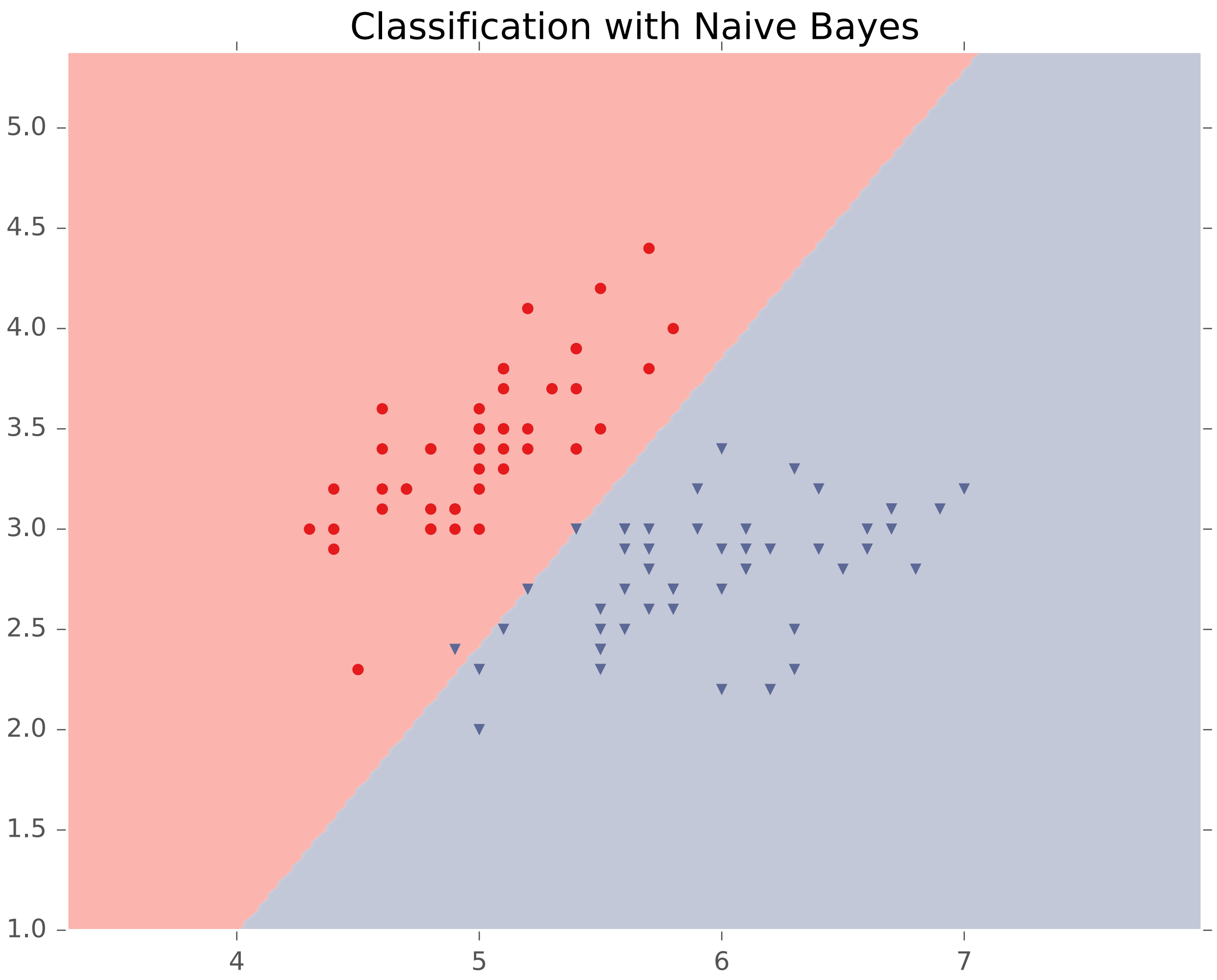


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

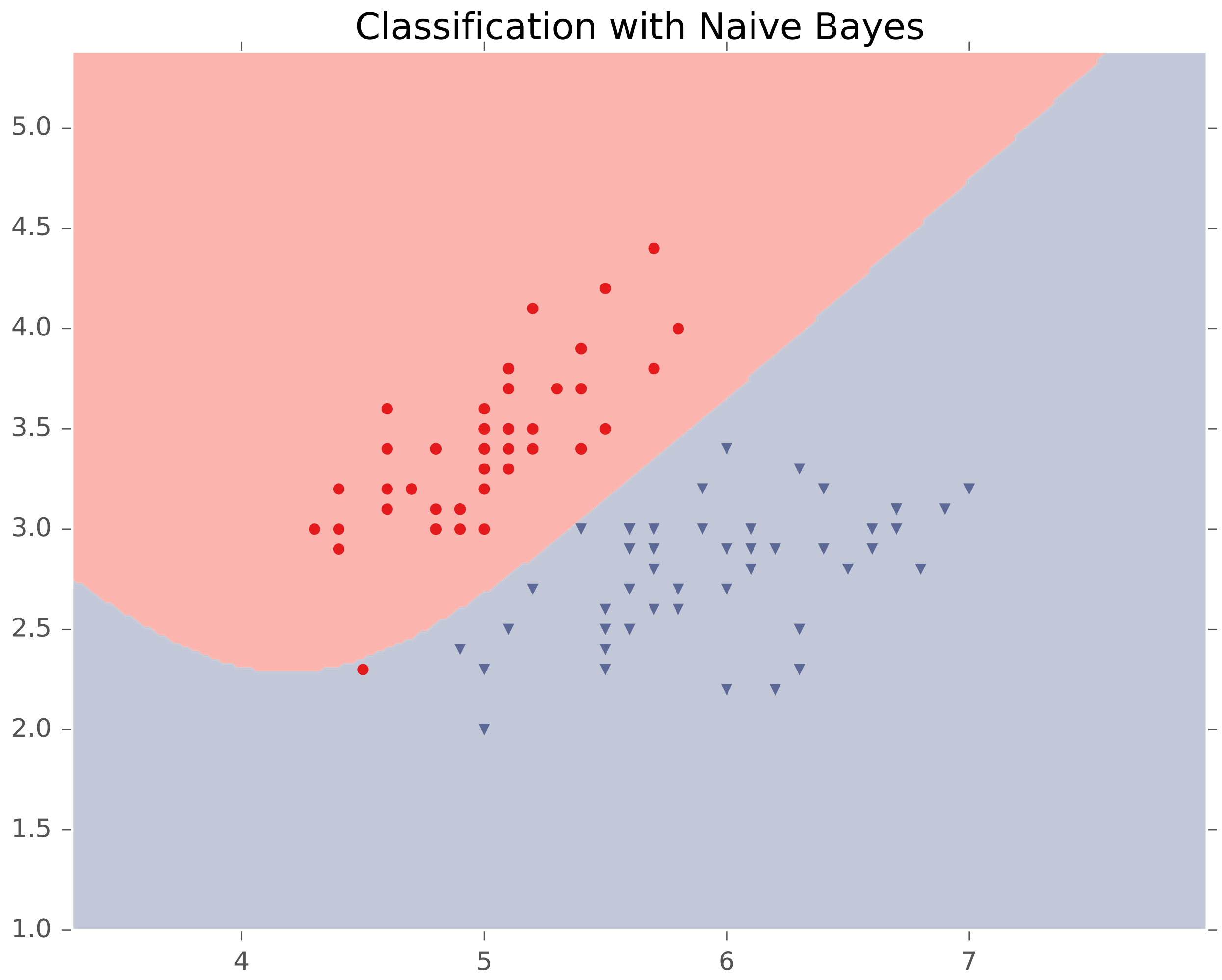


Figure courtesy of Matt Gormley