10-301/601: Introduction to Machine Learning Lecture 20 - Naïve Bayes

Henry Chai

7/12/23

Front Matter

- Announcements:
  - PA4 released 6/15, due 7/13 (tomorrow) at 11:59 PM
  - PA5 released 7/13 (tomorrow), due 7/20 at 11:59 PM
  - Quiz 7: Unsupervised Learning & Naïve Bayes on 7/18
- Recommended Readings:
  - Mitchell, draft chapter on Naïve Bayes & logistic regression
  - Murphy, Chapter 3.5

What is ~~Machine Learning~~ 10-301/601?

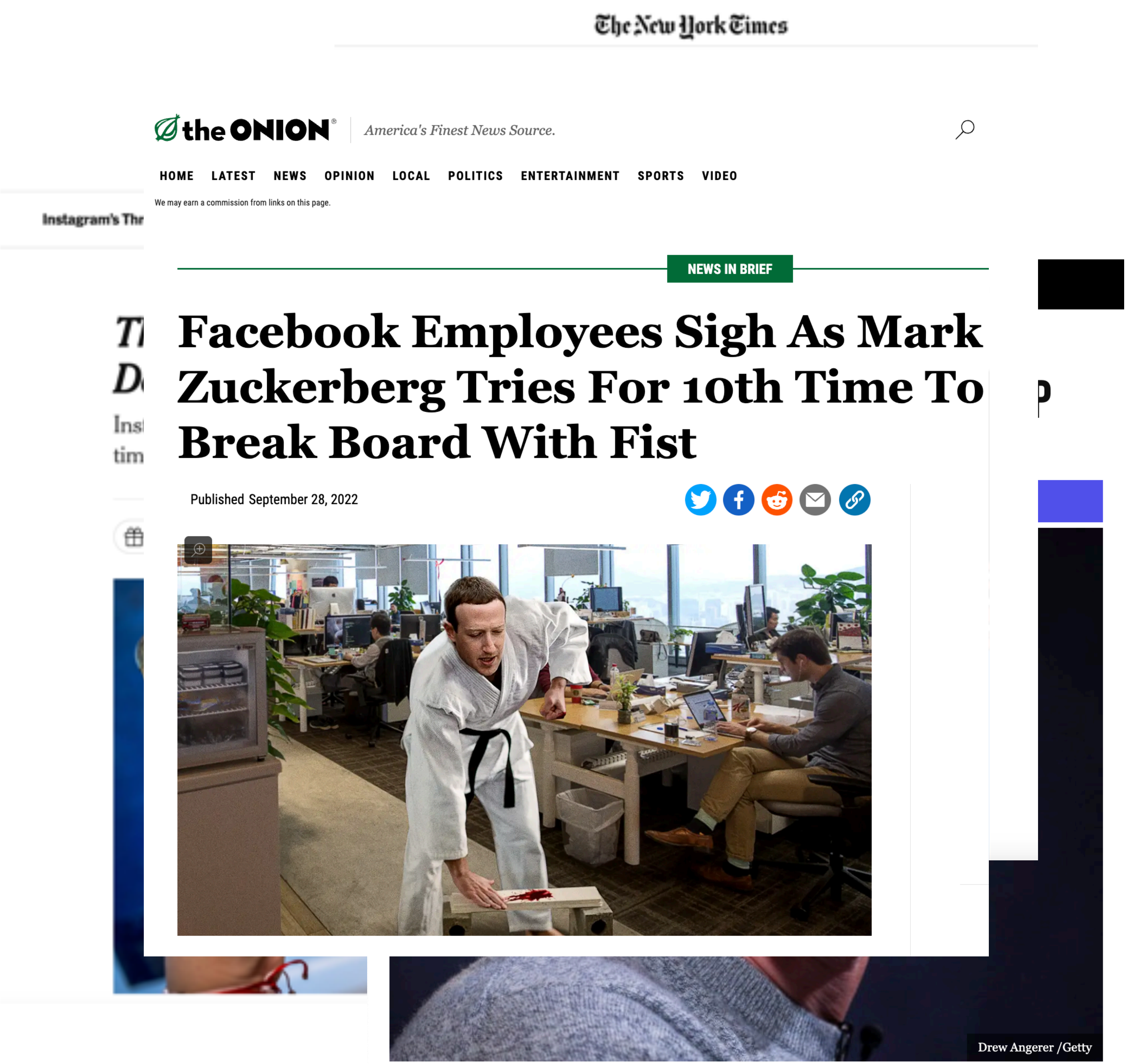
- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - SVMs
  - Linear Regression
  - Neural Networks
- Unsupervised Models
  - K-means
  - GMMs
  - PCA
- Graphical Models
  - Bayesian Networks
  - HMMs
- Learning Theory
- Reinforcement Learning
- Important Concepts
  - Feature Engineering and Kernels
  - Regularization and Overfitting
  - Experimental Design
  - Ensemble Methods

What is ~~Machine Learning~~ 10-301/601?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - SVMs
  - Linear Regression
  - Neural Networks
- Unsupervised Models
  - K-means
  - GMMs
  - PCA
- **Graphical Models**
  - Bayesian Networks
  - HMMs
- Learning Theory
- Reinforcement Learning
- Important Concepts
  - Feature Engineering and Kernels
  - Regularization and Overfitting
  - Experimental Design
  - Ensemble Methods

Text Data

- https://www.nytimes.com/2023/07/06/technology/threads-downloads-twitter.html?searchResultPosition=1
- https://www.breitbart.com/tech/2023/07/06/sanely-run-mark-zuckerbergs-twitter-clone-censors-donald-trump-jr-on-day-one/
- https://www.nytimes.com/2023/07/01/technology/elon-musk-mark-zuckerberg-cage-match.html?searchResultPosition=2
- https://www.theonion.com/facebook-employees-sigh-as-mark-zuckerberg-tries-for-10-1849518797



Text Data


The TOOTH Book
Dr. Seuss
POCKET ROCKS
SHEREE FITCH
One fish two fish red fish blue fish
HOP ON POP
By Dr. Seuss
The SIMPLEST SEUSS for YOUNGEST USE
THE CAT IN THE HAT
By Dr. Seuss
See, Touch, Feel
A first sensory book to share with your baby
Lois Ehlert
My Mother?
by P.D. Eastman
Go, Dog. Go!
by P.D. Eastman
BLUE FISH
RED FISH
NEVER touch a dinosaur!
Chicka Chicka Boom Boom
by Bill Martin Jr and John Archambault
illustrated by Lois Ehlert
The CRAYONS' BOOK of FEELINGS
幼儿启蒙大全
YOU ER QI MENG DA QUAN
Big Dog... Little Dog
by P. D. Eastman
FRIENDS COLLECTION
Pooh & Eeyore

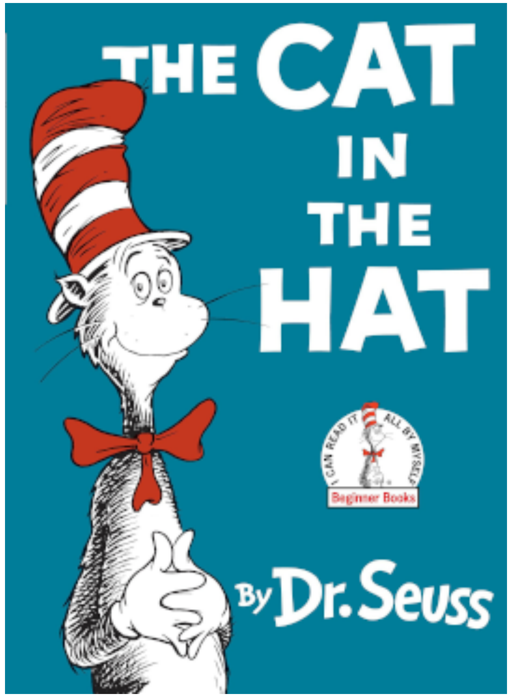
Bag-of-Words Model

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)

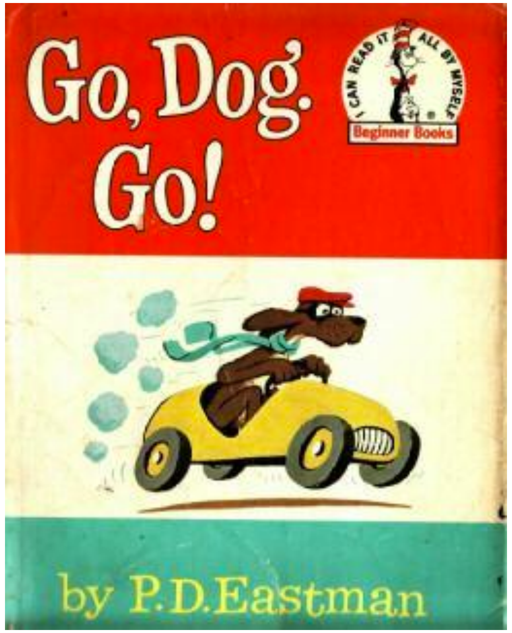


Source: https://en.wikipedia.org/wiki/The_Cat_in_the_Hat#/media/File:The_Cat_in_the_Hat.png

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)

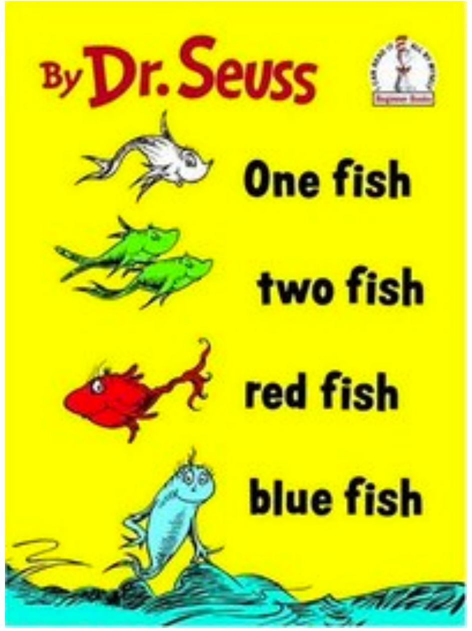


Source: https://en.wikipedia.org/wiki/Go,_Dog._Go!#/media/File:Go_Dog_Go.jpg

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)

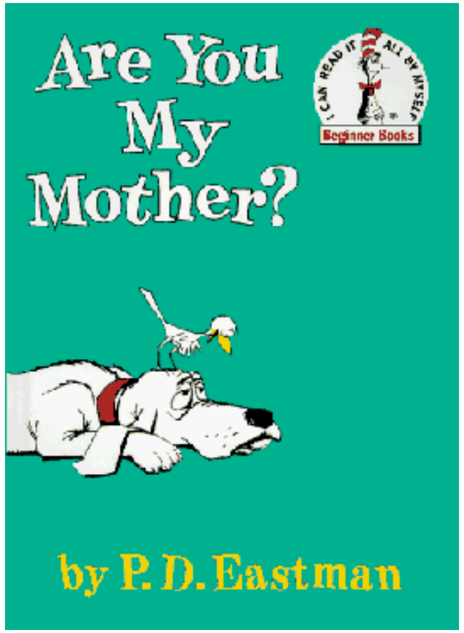


Source: https://en.wikipedia.org/wiki/One_Fish,_Two_Fish,_Red_Fish,_Blue_Fish#/media/File:One_Fish_Two_Fish_Red_Fish_Blue_Fish_(cover_art).jpg

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

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 (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 (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

When survey is active, respond at **pollev.com/301601polls**

Lecture 20 Polls

0 done

0 underway

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Given 6 binary features $\mathbf{x} = [x_1, ..., x_6]^T$ and a binary label y, how many parameters are needed to fully specify the distribution $P(\mathbf{x}|Y = y)$?

- $2^6 = 64$
- $2^6 - 1 = 63$
- $2(2^6) = 128$
- $2(2^6 - 1) = 126$

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

$2(2^6 - 1) = 126$

Naïve Bayes Assumption

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

$$P(X|Y) = P(X_1 \cap X_2 \cap \ldots X_D | Y) = \prod_{d=1}^{D} P(X_d | Y)$$

- Pros:
  - Significantly reduces computational complexity
  - Also can combat overfitting by reducing model complexity
- Cons:
  - This is a terrible assumption!
    - We can relax this assumption via Bayesian Networks (Monday!)

Given 6 binary features $\mathbf{x} = [x_1, \ldots, x_6]^T$ and a binary label y, how many parameters are needed to fully specify the distribution $P(\mathbf{x}|Y = y)$ with the naïve Bayes assumption?

you don't need to learn say both $P(X_d=1|Y=1)$ and $P(X_d=0|Y=1)$

$P(X_d=1|Y=1)$ and $P(X_d=1|Y=0)$ $\forall d$

- 6
- $6 - 1 = 5$
- $2(6) = 12$
- $2(6-1) = 10$

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 Naive Bayes assumption of conditional independence
  - Assume iid data points
  - Parameters (Example): $\pi = P(Y=1)$, $\theta_{d,y} = P(X_d = 1 \mid Y = y)$
- 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 equal 0 and solve

$$\pi = P(Y=1)$$

Setting the Parameters via MLE

Ex: $\dfrac{\partial \ell}{\partial \theta_{2,0}}$

$$\ell_D(\pi, \theta) = \log P(D = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N} \mid \pi, \theta)$$

$$= \log \prod_{n=1}^{N} P(x^{(n)}, y^{(n)} \mid \pi, \theta) = \sum_{n=1}^{N} \log P(x^{(n)}, y^{(n)} \mid \pi, \theta)$$

$$= \sum_{n=1}^{N} \log \underbrace{P(x^{(n)} \mid y^{(n)}, \pi, \theta)} \underbrace{P(y^{(n)} \mid \pi, \theta)}$$

N.B. assumption

$$= \sum_{n=1}^{N} \log \left(\prod_{d=1}^{D} P(x_d^{(n)} \mid y^{(n)}, \theta) \right) P(y^{(n)} \mid \pi)$$

$$\theta = [\theta_{1,0}, \theta_{1,1}, \ldots]$$

$$= \sum_{n=1}^{N} \left(\left(\sum_{d=1}^{D} \log P(x_d^{(n)} \mid y^{(n)}, \theta) \right) + \log P(y^{(n)} \mid \pi) \right)$$

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

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

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 $X_d = 1$

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)
  - $X_d | Y = y \sim \text{Categorical}\left(\theta_{d,1,y}, \ldots, \theta_{d,K-1,y}\right)$
  - $\hat{\theta}_{d,k,y} = {}^{N_{Y=y,\, X_d=k}}/_{N_{Y=y}}$
    - $N_{Y=y}$ = # 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
  - $X_d | Y = y \sim \text{Gaussian}\left(\mu_{d,y}, \sigma_{d,y}^2\right)$
  - $\hat{\mu}_{d,y} = \frac{1}{N_{Y=y}} \sum_{n: y^{(n)} = y} x_d^{(n)}$
  - $\hat{\sigma}_{d,y}^2 = \frac{1}{N_{Y=y}} \sum_{n: y^{(n)} = y} \left(x_d^{(n)} - \hat{\mu}_{d,y}\right)^2$
    - $N_{Y=y}$ = # of data points with label y

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

Visualizing Gaussian Naïve Bayes (2 classes)

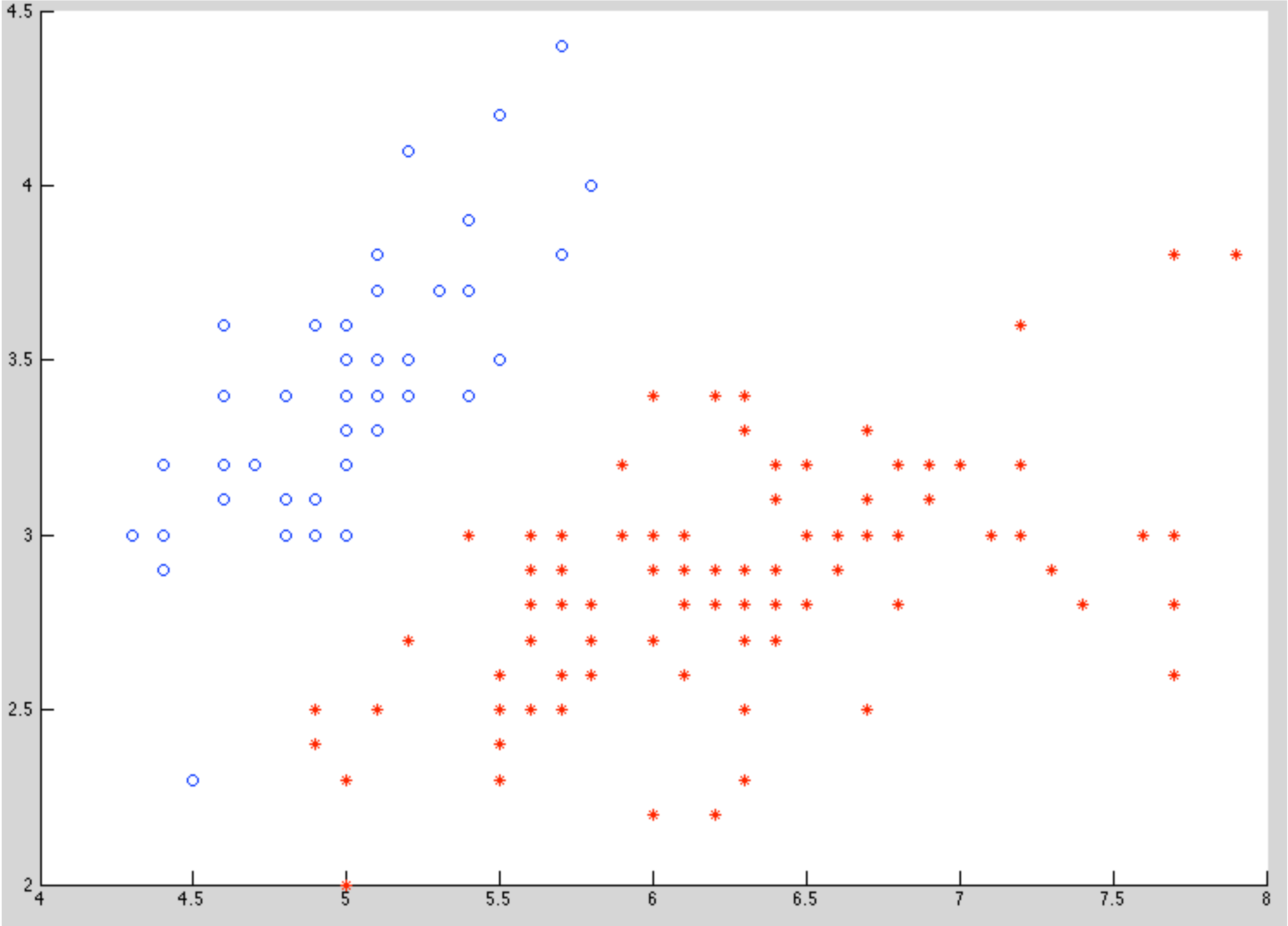


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes)

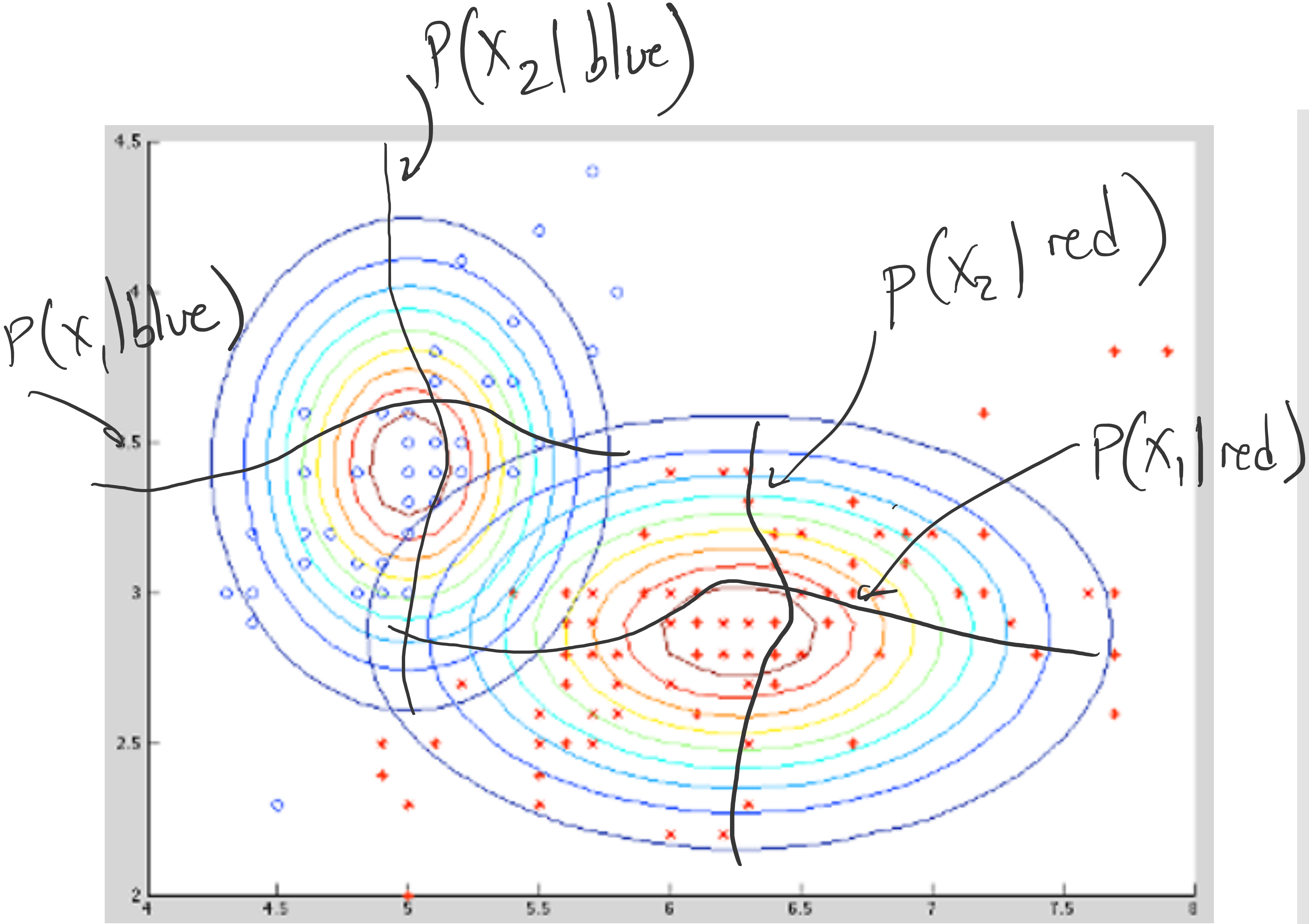


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes, equal variances)

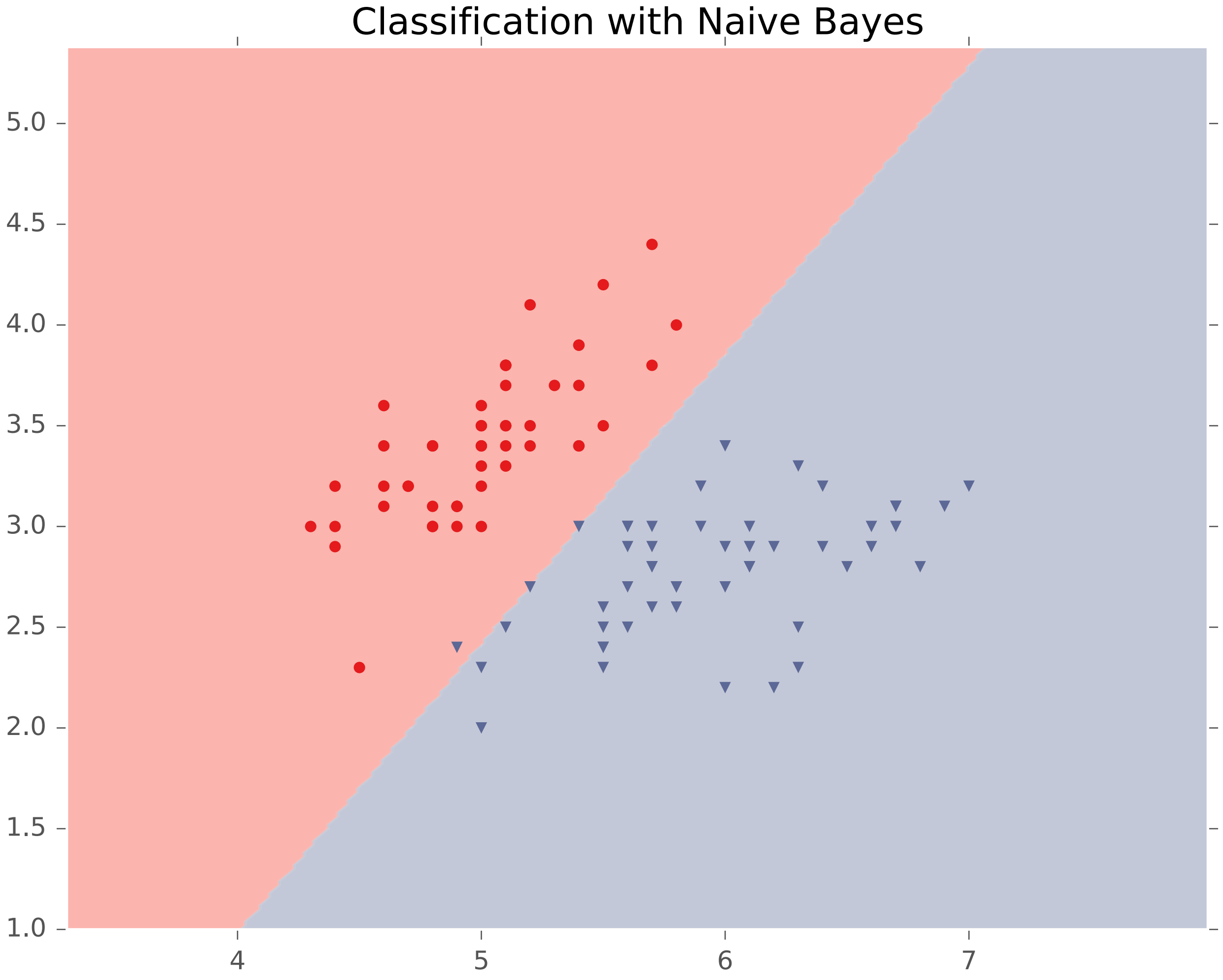


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

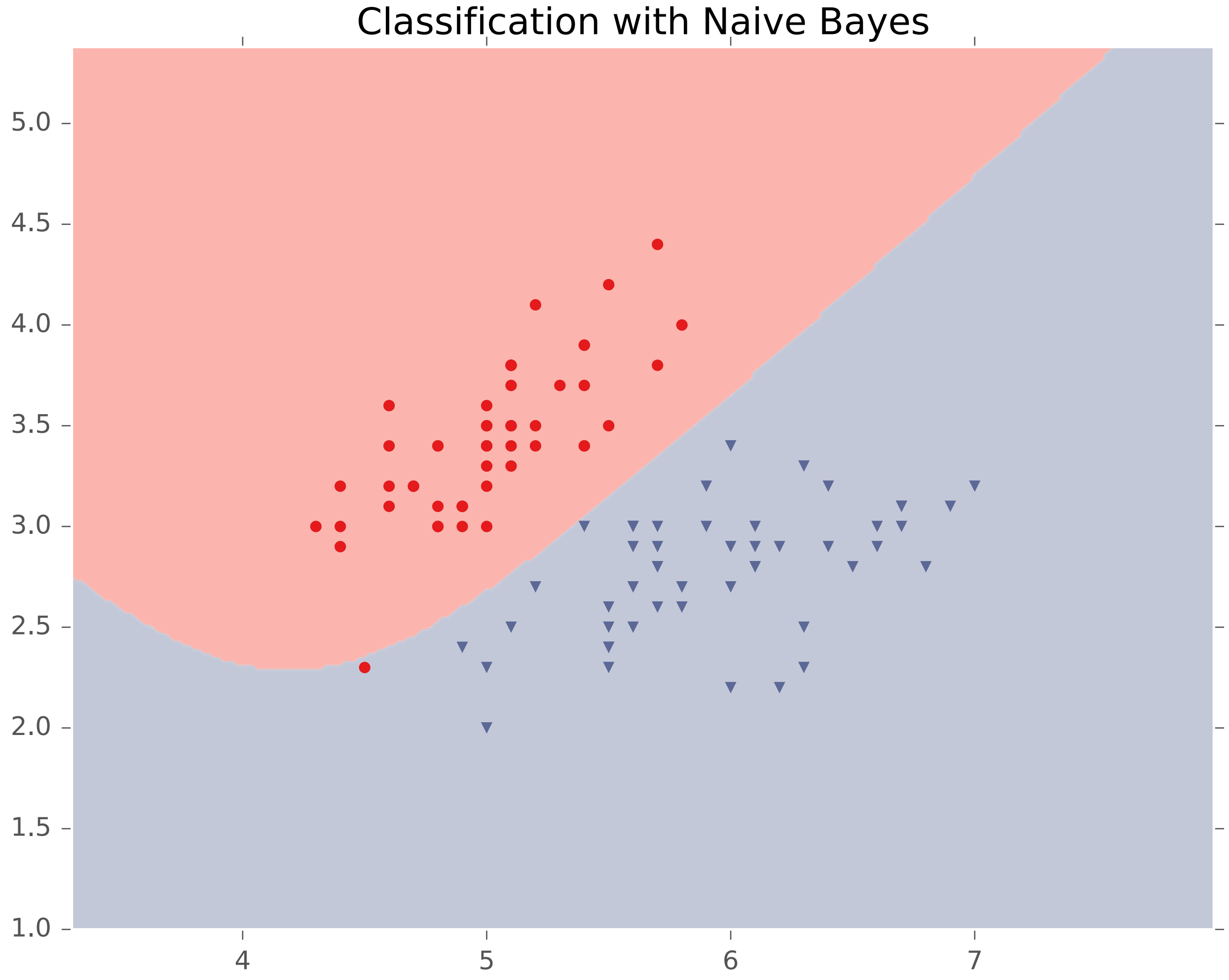


Figure courtesy of Matt Gormley

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 $X_d = 1$

Bernoulli Naïve Bayes: Making Predictions

- Given a test data point $\boldsymbol{x}' = [x_1', \dots, x_D']^T$

$$P(Y=1 \mid x') \propto P(x' \mid Y=1)\,P(Y=1)$$

$$= \hat{\pi} \prod_{d=1}^{D} P(x_d' \mid Y=1)$$

$$= \hat{\pi} \prod_{d=1}^{D} \hat{\theta}_{d,1}^{\,x_d'} (1-\hat{\theta}_{d,1})^{1-x_d'}$$

$$P(Y=0 \mid x') = (1-\hat{\pi}) \prod_{d=1}^{D} \hat{\theta}_{d,0}^{\,x_d'} (1-\hat{\theta}_{d,0})^{1-x_d'}$$

$$\hat{y} = \begin{cases} 1 & \text{if } P(Y=1 \mid x') > P(Y=0 \mid x') \\ 0 & \text{otherwise} \end{cases}$$

What if some Word-Label pair never appears in our training data?

Bernoulli Naive Bayes: Making Predictions

- Given a test data point $\boldsymbol{x}' = [x_1', \dots, x_D']^T$

What if some Word-Label pair never appears in our training data?

- Given a test data point $\boldsymbol{x}' = [x'_1, \dots, x'_D]^T$

$$P(Y = 1|\boldsymbol{x}') \propto P(Y = 1)P(\boldsymbol{x}'|Y = 1)$$
$$= \hat{\pi} \prod_{d=1}^{D} \hat{\theta}_{d,1}^{x'_d} \left(1 - \hat{\theta}_{d,1}\right)^{1-x'_d}$$

$$P(Y = 0|\boldsymbol{x}') \propto (1 - \hat{\pi}) \prod_{d=1}^{D} \hat{\theta}_{d,0}^{x'_d} \left(1 - \hat{\theta}_{d,0}\right)^{1-x'_d}$$

$$\hat{y} = \begin{cases} 1 \text{ if } \hat{\pi} \prod_{d=1}^{D} \hat{\theta}_{d,1}^{x'_d} \left(1 - \hat{\theta}_{d,1}\right)^{1-x'_d} > \\ \quad (1 - \hat{\pi}) \prod_{d=1}^{D} \hat{\theta}_{d,0}^{x'_d} \left(1 - \hat{\theta}_{d,0}\right)^{1-x'_d} \\ 0 \text{ otherwise} \end{cases}$$

What if some Word-Label pair never appears in our training data?

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

The Cat in the Hat gets a Dog (by ???)

- If some $\hat{\theta}_{d,y} = 0$ and that word appears in our test data $\boldsymbol{x}'$, then $P(Y = y|\boldsymbol{x}') = 0$ even if all the other features in $\boldsymbol{x}'$ point to the label being y!
- The model has been overfit to the training data…
- We can address this with a prior over the parameters!

Setting the Parameters via MAP

- 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
  - $X_d | Y = y \sim \text{Bernoulli}(\theta_{d,y})$ and $\underline{\theta_{d,y} \sim \text{Beta}(\alpha, \beta)}$
  - $\hat{\theta}_{d,y} = {}^{N_{Y=y,\, X_d=1} + (\alpha-1)}\!/_{N_{Y=y} + (\alpha-1) + (\beta-1)}$
    - $N_{Y=y}$ = # of data points with label y
    - $N_{Y=y,\, X_d=1}$ = # of data points with label y and feature $X_d = 1$
    - α and β are "pseudocounts" of imagined data points that help avoid zero-probability predictions.
    - Common choice: $\alpha = \beta = 2$

What can we do when this is a bad/incorrect assumption, e.g., when our features are words in a sentence?

- 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

Key Takeaways

- Text data
    - Bag-of-words feature representation
- Naïve Bayes
    - Conditional independence assumption
        - Pros and cons
    - Different Naïve Bayes models based on type of features
    - MLE vs. MAP for Bernoulli Naïve Bayes

Naïve Bayes Case Study: Categorizing News Articles (Mitchell, Chapter 6)

- 1000 Usenet articles from 20 different newsgroups:

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

- Multinomial Naïve Bayes classifier using bag-of-word features (word frequencies) achieves 89% test accuracy!