10-701: Introduction to Machine Learning

Lecture 6 –MAP & Naïve Bayes

Hoda Heidari

* Slides adopted from F24 offering of 10701 by Henry Chai.

Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \to \mathcal{Y}$
  - Classifier, $h : \mathcal{X} \to \mathcal{Y}$
  - Goal: find a classifier, h, that best approximates c^*
- Now:
  - (Unknown) Target *distribution*, $y \sim p^*(Y|\boldsymbol{x})$
  - Distribution, $p(Y|\boldsymbol{x})$
  - Goal: find a distribution, p, that best approximates p^*
  - Suppose p comes from a parametric family of distributions, parameterized by θ

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 $\hat{\theta} = \underset{\theta}{\text{argmax}}\ p(\mathcal{D}|\theta)$
  - MAP finds $\hat{\theta} = \underset{\theta}{\text{argmax}}\ p(\theta|\mathcal{D})$

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 $\hat{\theta} = \underset{\theta}{\text{argmax}}\ p(\mathcal{D}|\theta)$
  - MAP finds $\hat{\theta} = \underset{\theta}{\text{argmax}}\ p(\theta|\mathcal{D})$

$$= \underset{\theta}{\text{argmax}}\ p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$$

$$= \underset{\theta}{\text{argmax}}\ p(\mathcal{D}|\theta)p(\theta)$$

likelihood prior

$$= \underset{\theta}{\text{argmax}}\ \log p(\mathcal{D}|\theta) + \log p(\theta)$$

log-posterior

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

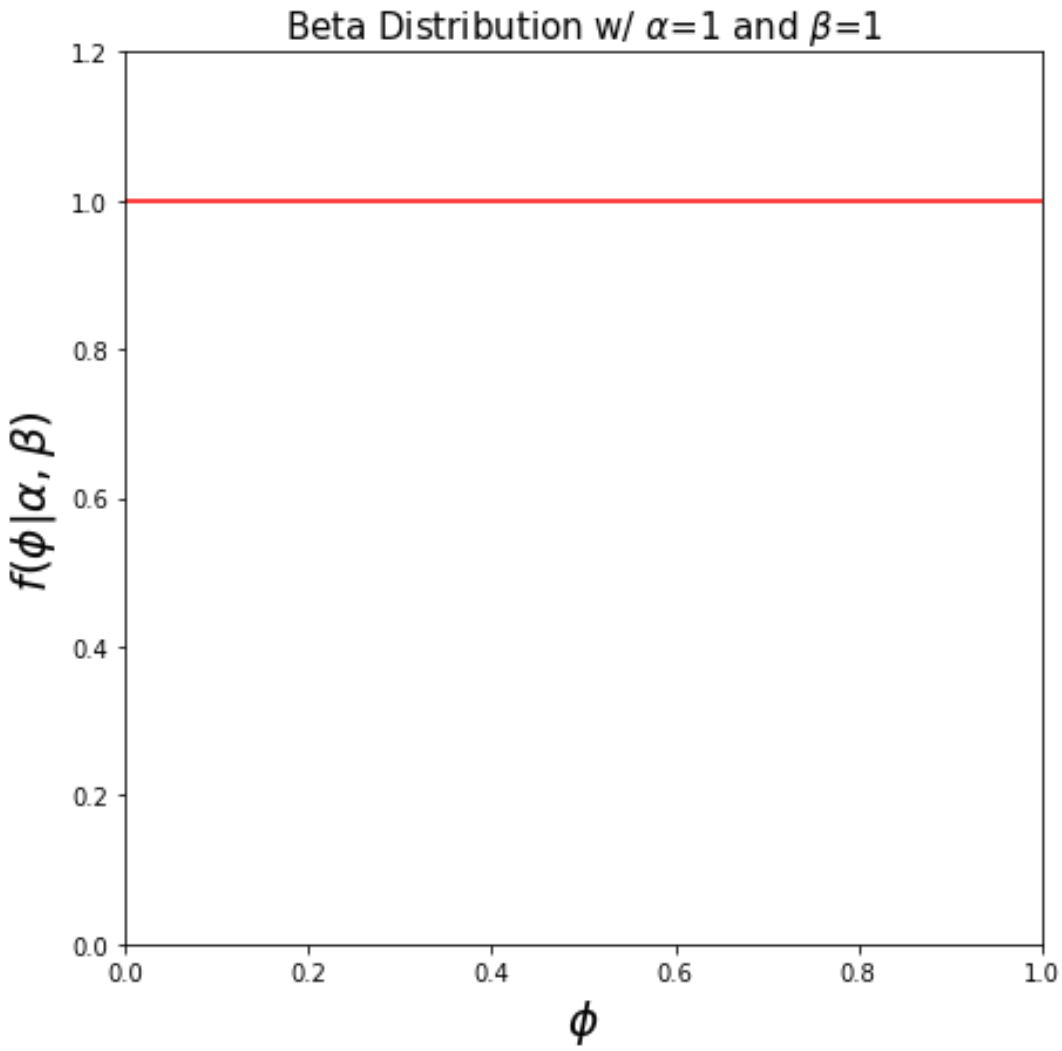
$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

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

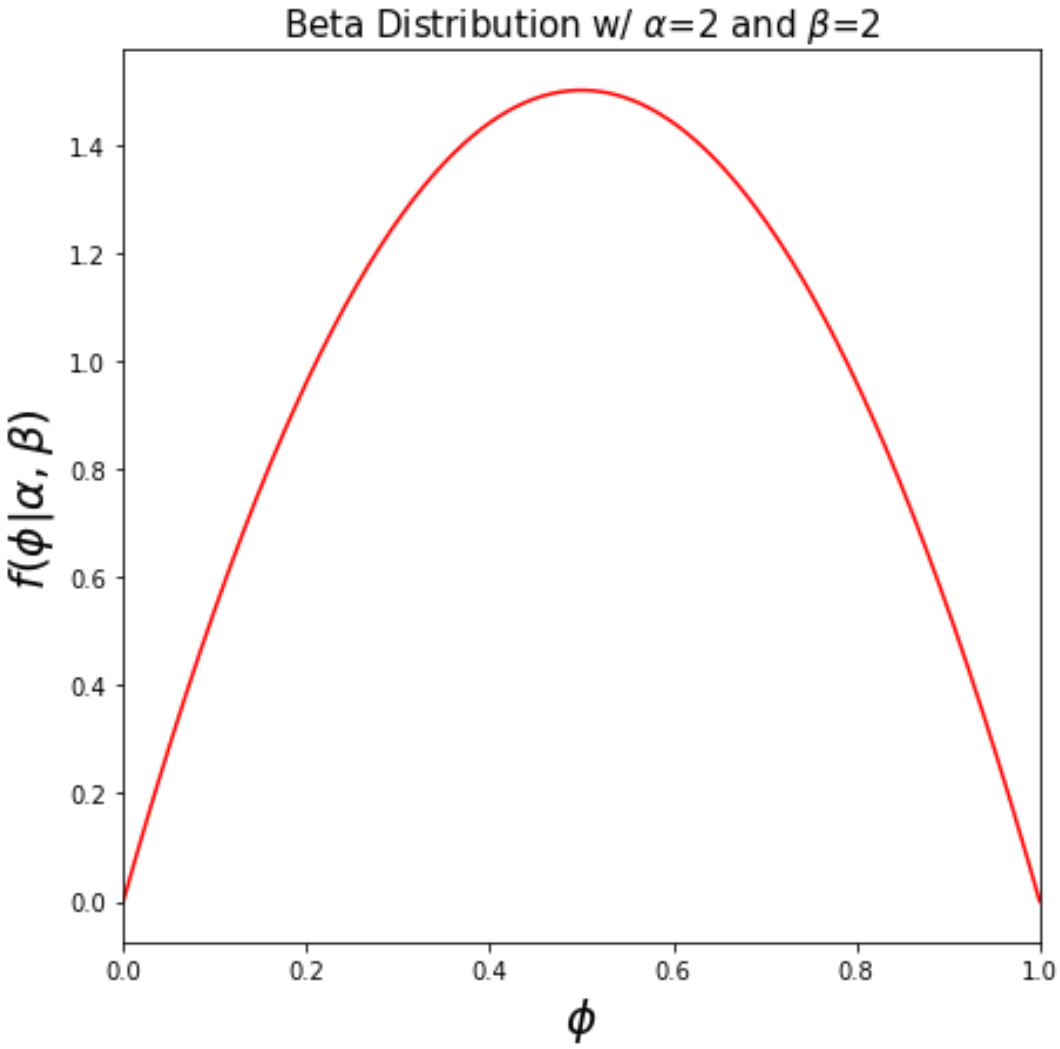
$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

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

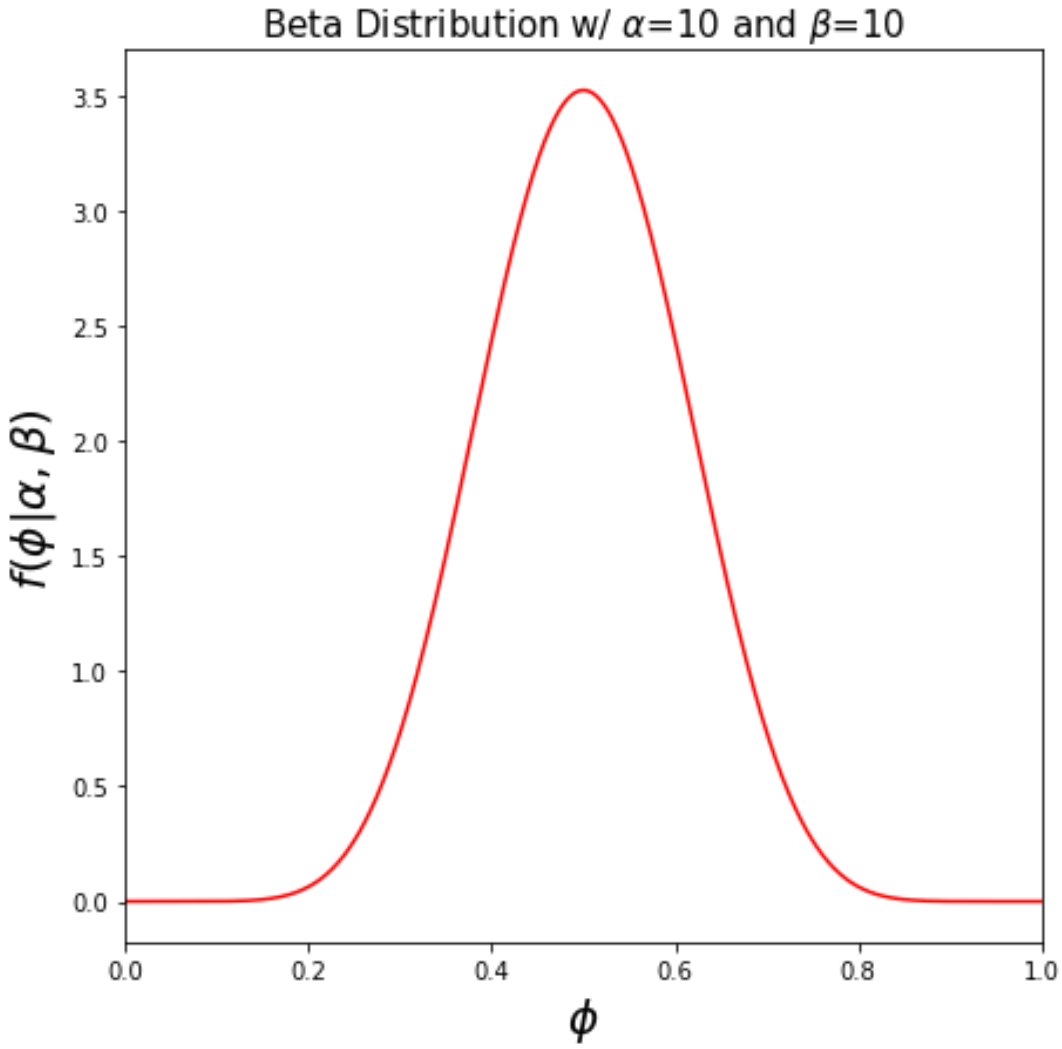
Beta Distribution



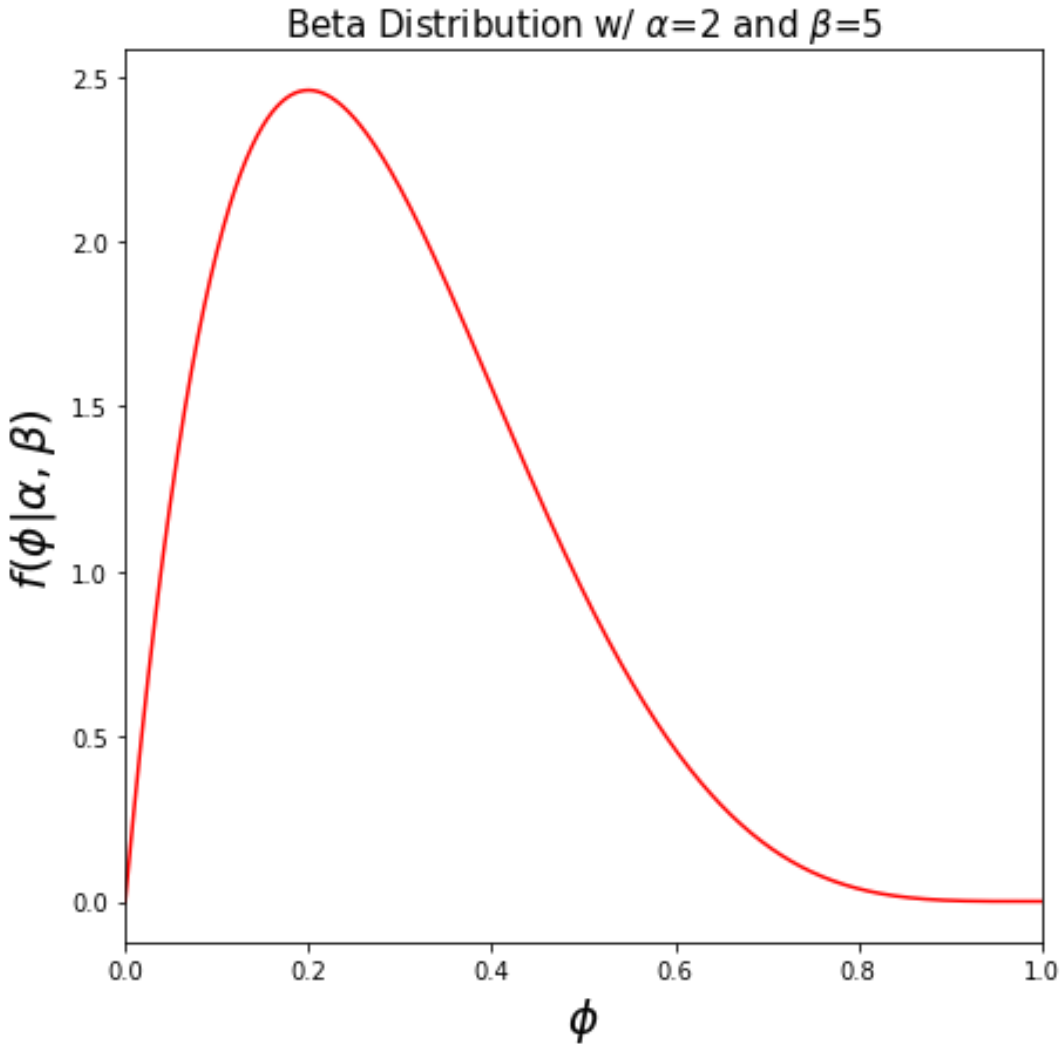
Beta Distribution



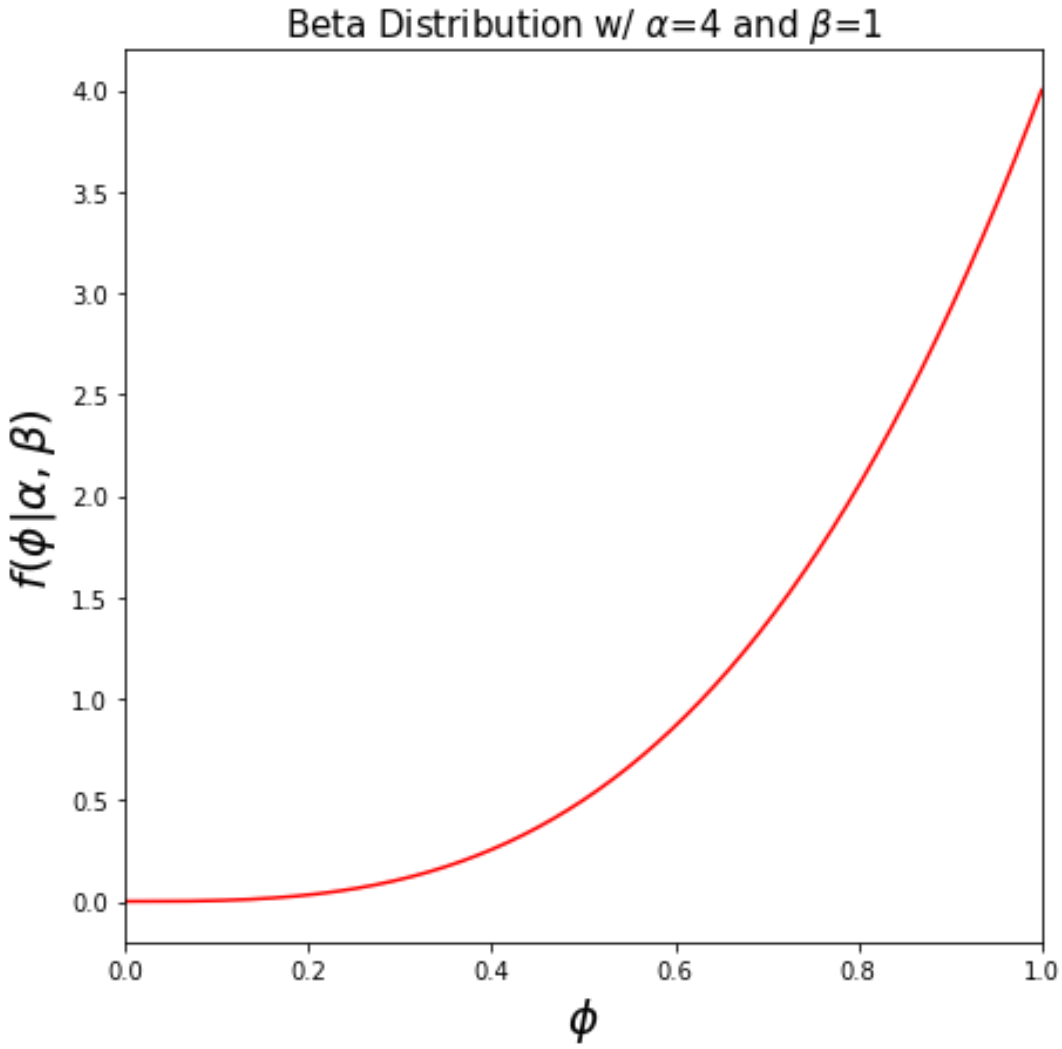
Beta Distribution



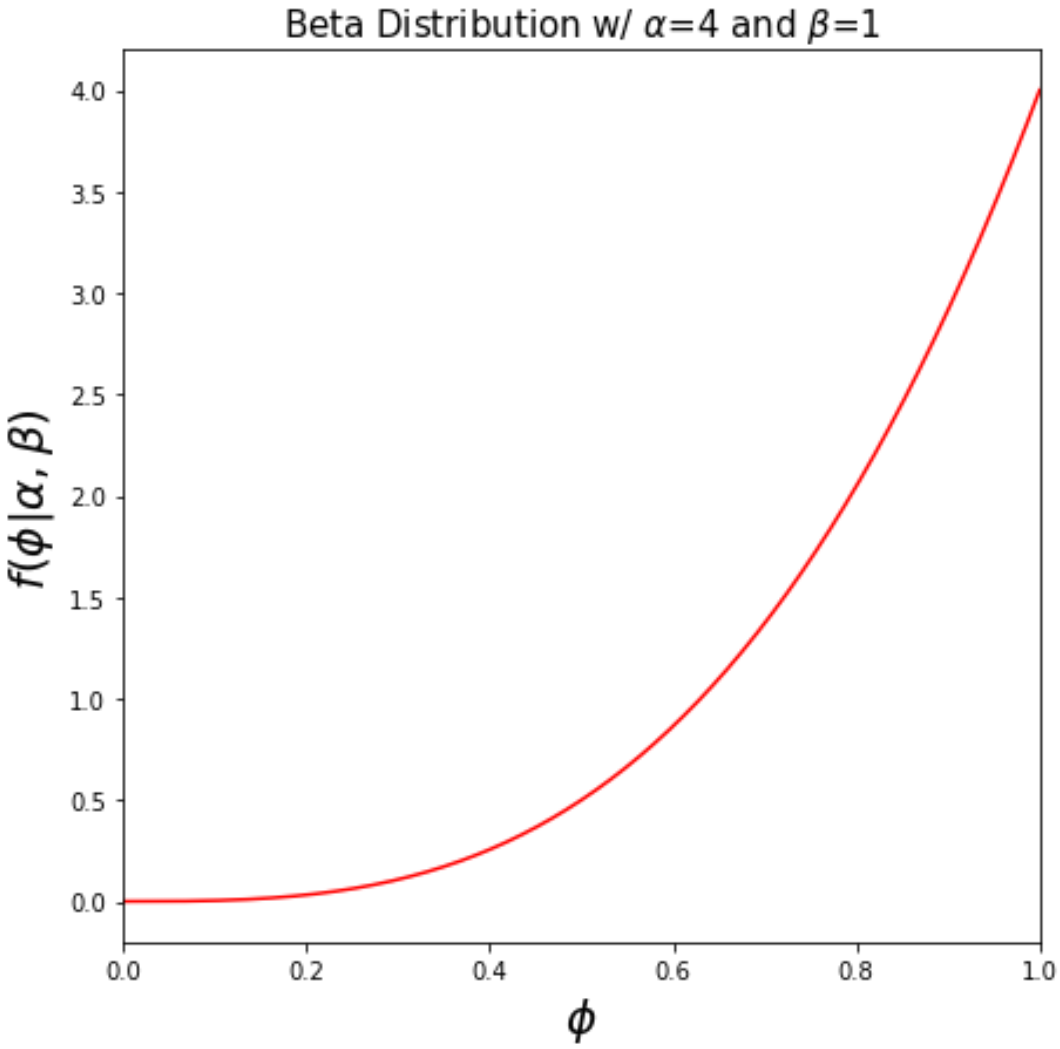
Beta Distribution



Beta Distribution



Okay, but why should we use this strange distribution as a prior?



Conjugate Priors

- For a given likelihood function $p(\mathcal{D}|\theta)$, a prior $p(\theta)$ is called a *conjugate prior* if the resulting posterior distribution $p(\theta|\mathcal{D})$ is in the same family as $p(\theta)$ i.e., $p(\theta|\mathcal{D})$ and $p(\theta)$ are the same type of random variable just with different parameters
  - We like conjugate priors because they are mathematically convenient
  - However, we do not **have** to use a conjugate prior if it doesn't align with our actual prior belief.

Example: Beta-Binomial Conjugacy

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

$$p(x|\alpha,\beta) =$$

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

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

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

Example: Beta-Binomial Conjugacy

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

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

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

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

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

Beta-Binomial MAP

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

Beta-Binomial MAP

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

$$\ell(\phi) = \log f\left(\phi \middle| (\alpha + x^{(1)} + x^{(2)} + \cdots x^{(N)}\right),$$

$$\left(\beta + \left(1 - x^{(1)}\right) + \left(1 - x^{(2)}\right) + \cdots + \left(1 - x^{(N)}\right)\right)$$

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

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

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

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

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

$$\phi_{MAP} =$$

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

$$\phi_{MAP} =$$

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

$$\phi_{MAP} =$$

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

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

Key Takeaways

- Two ways of estimating the parameters of a probability distribution given samples of a random variable:
  - Maximum likelihood estimation – maximize the (log-)likelihood of the observations
  - Maximum a posteriori estimation – maximize the (log-)posterior of the parameters conditioned on the observations
    - Requires a prior distribution, drawn from background knowledge or domain expertise

Text Data


The TOOTH Book
POCKET ROCKS
THE CAT IN THE HAT
by Dr. Seuss
HOP ON POP
By Dr. Seuss
The SIMPLEST SEUSS for YOUNGEST USE
See, Touch, Feel
A first sensory book to share with your baby
Go, Dog. Go!
by P.D.Eastman
Are You My Mother?
by P.D. Eastman
Chicka Chicka Boom Boom
by Bill Martin Jr. and John Archambault
illustrated by Lois Ehlert
The CRAYONS' Book of FEELINGS
Big Dog... Little Dog...
by P. D. Eastman
幼儿启蒙大全

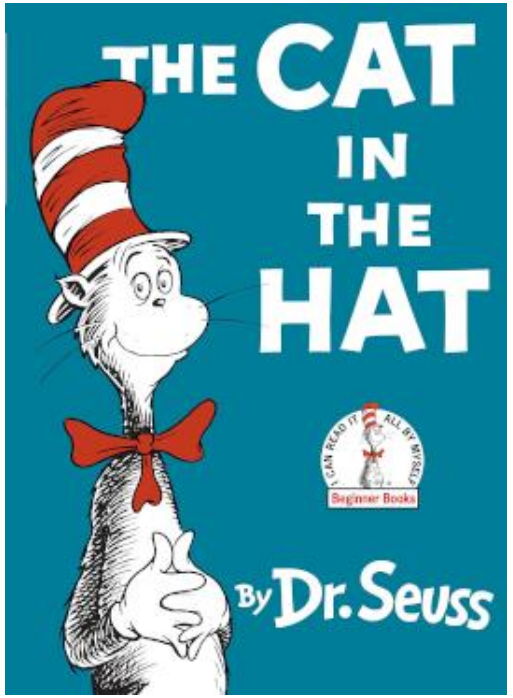
Bag-of-Words Model

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)

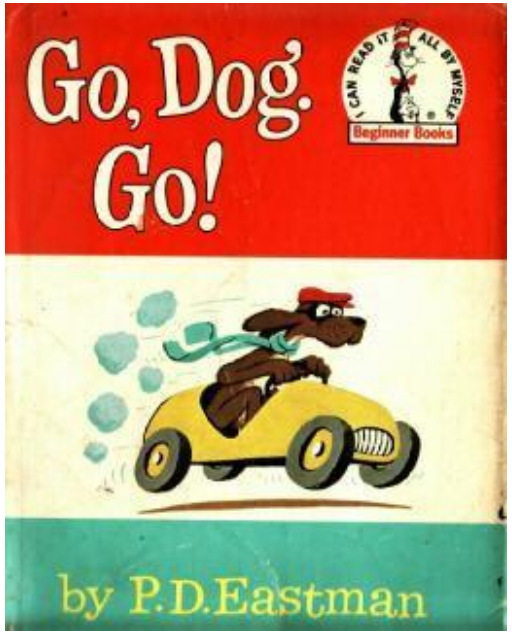


Source: https://en.wikipedia.org/wiki/The_Cat_in_the_Hat#/media/File:The_Cat_in_the_Hat.png

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)

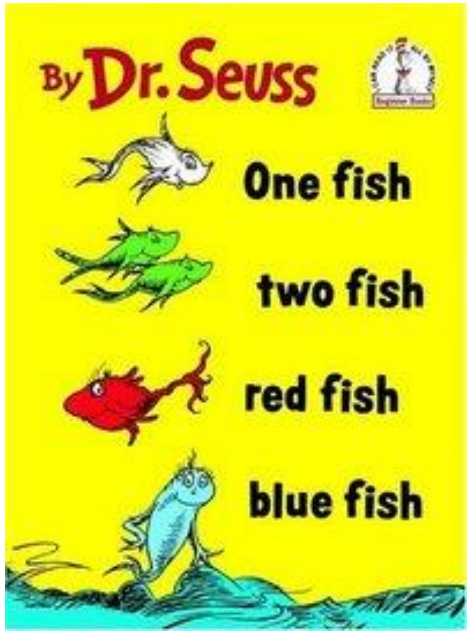


Source: https://en.wikipedia.org/wiki/Go,_Dog._Go!#/media/File:Go_Dog_Go.jpg

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)

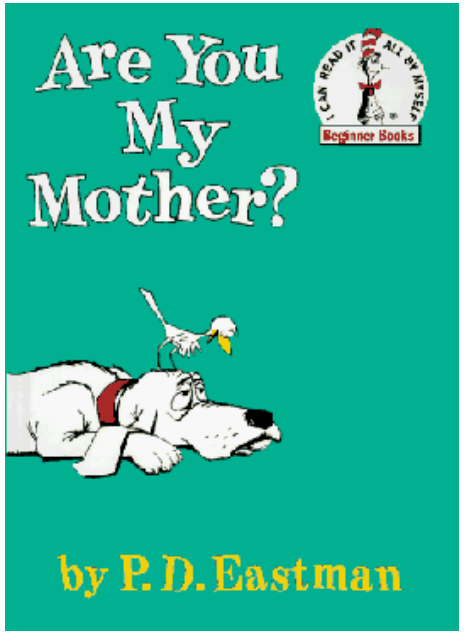


Source: https://en.wikipedia.org/wiki/One_Fish,_Two_Fish,_Red_Fish,_Blue_Fish#/media/File:One_Fish_Two_Fish_Red_Fish_Blue_Fish_(cover_art).jpg

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		
1	0	0	0	0	0		
1	1	0	0	0	0		
1	0	1	0	0	0		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

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	$1-\sum_{i=1}^{63}\theta_i$	$1-\sum_{i=64}^{127}\theta_i$

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

Recipe for Naïve Bayes

- Define a model space 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

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

Setting the Parameters via MLE

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

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

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

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

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

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

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

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)

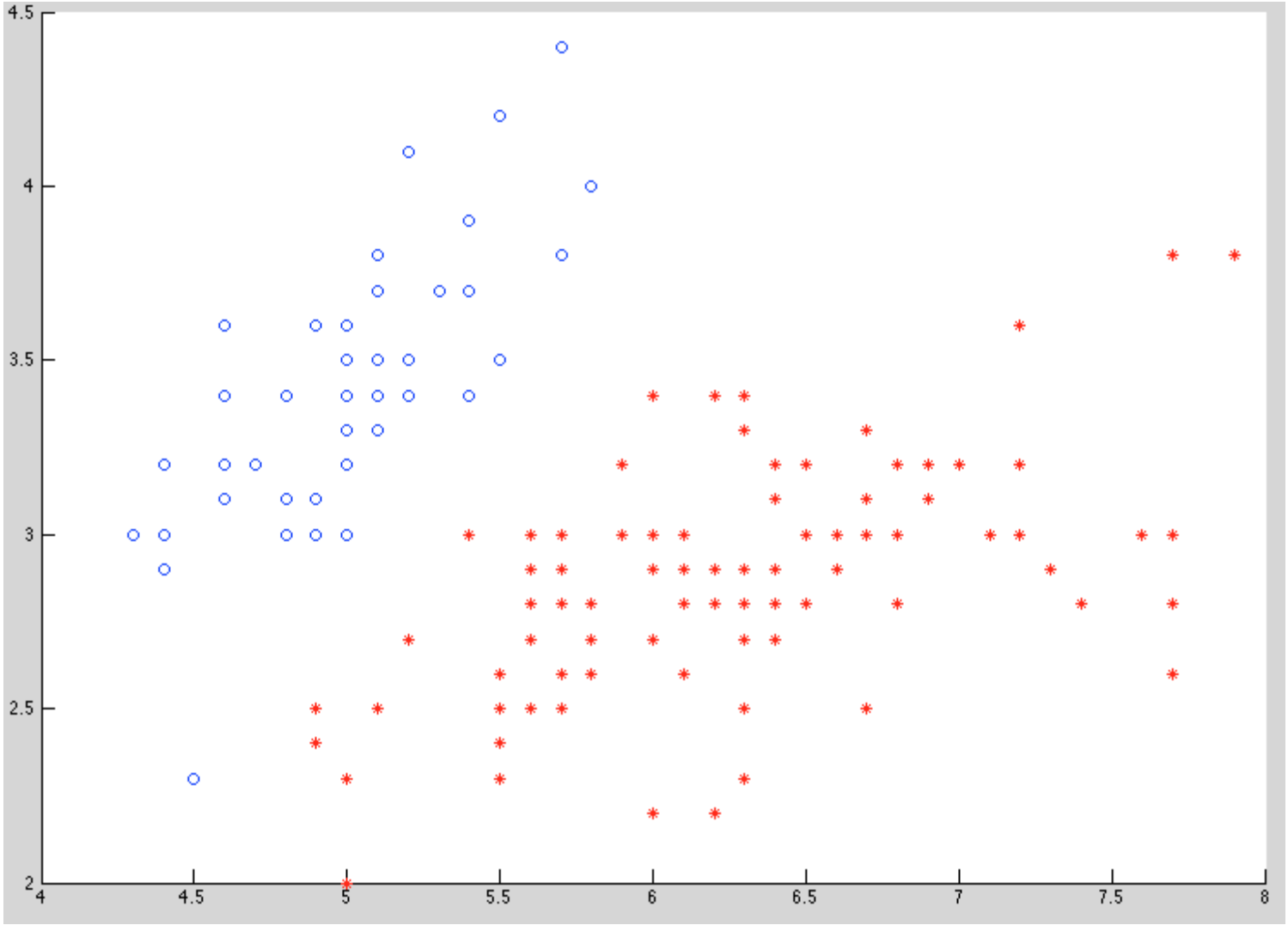


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes)

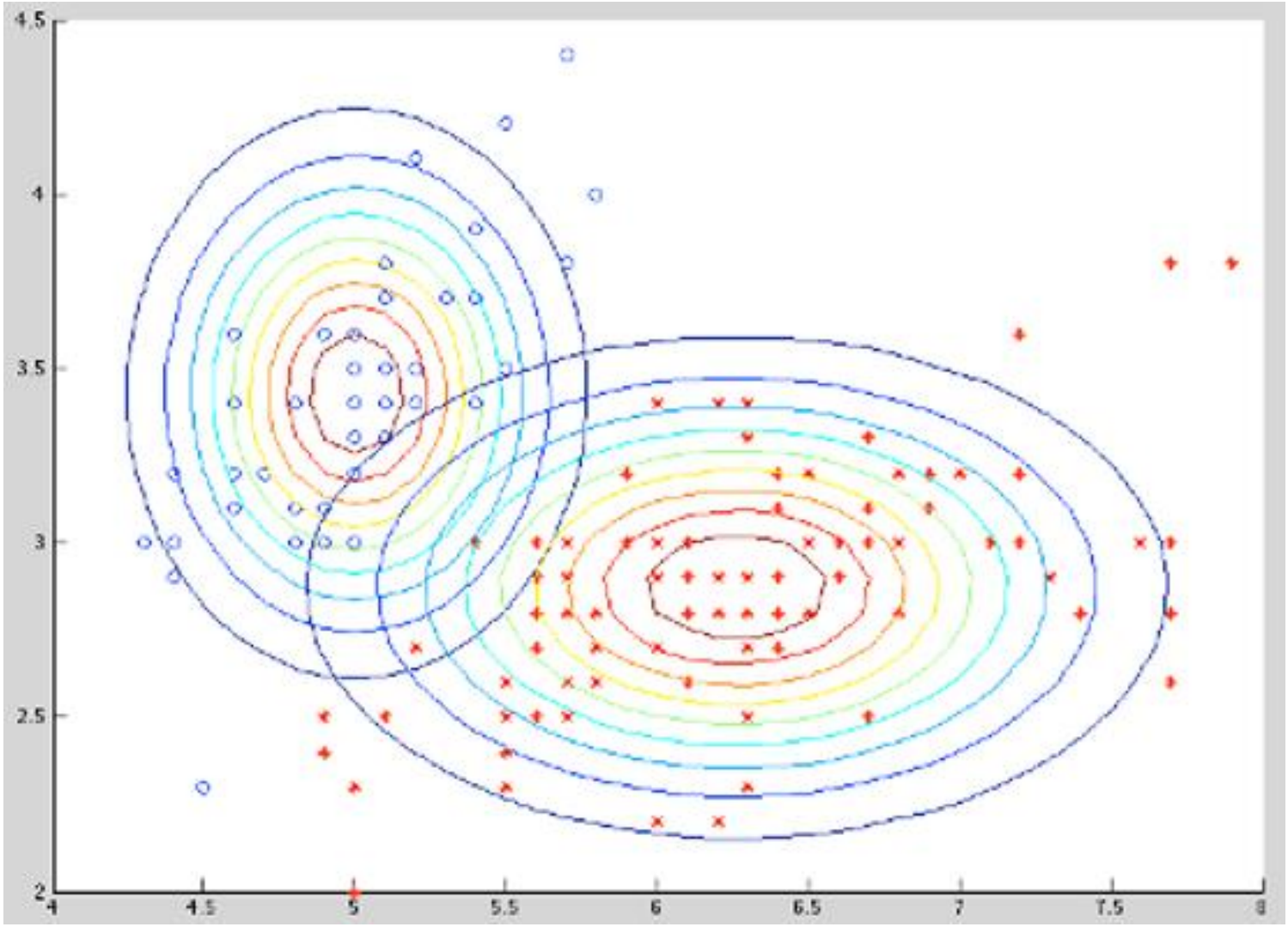


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes, equal variances)

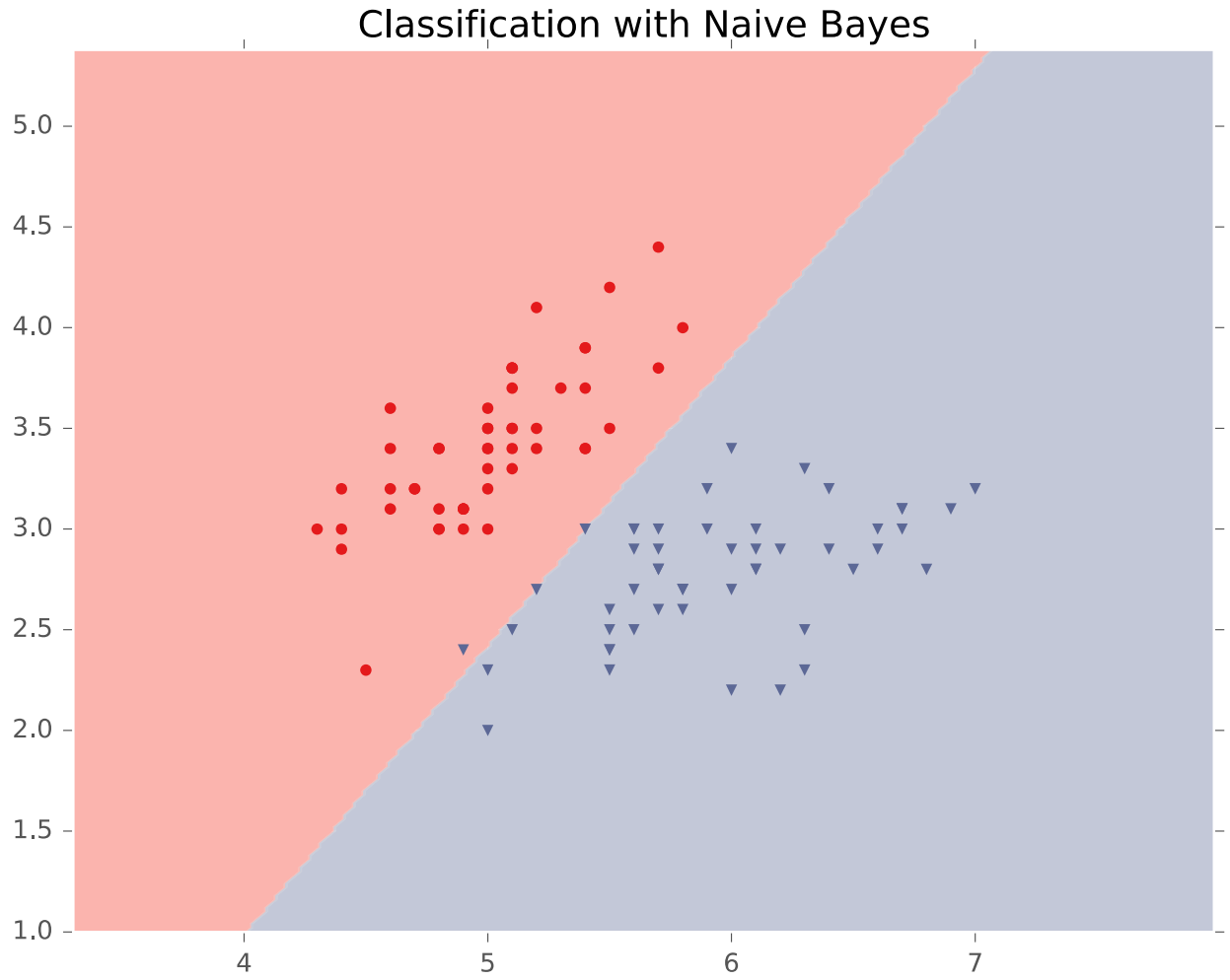


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

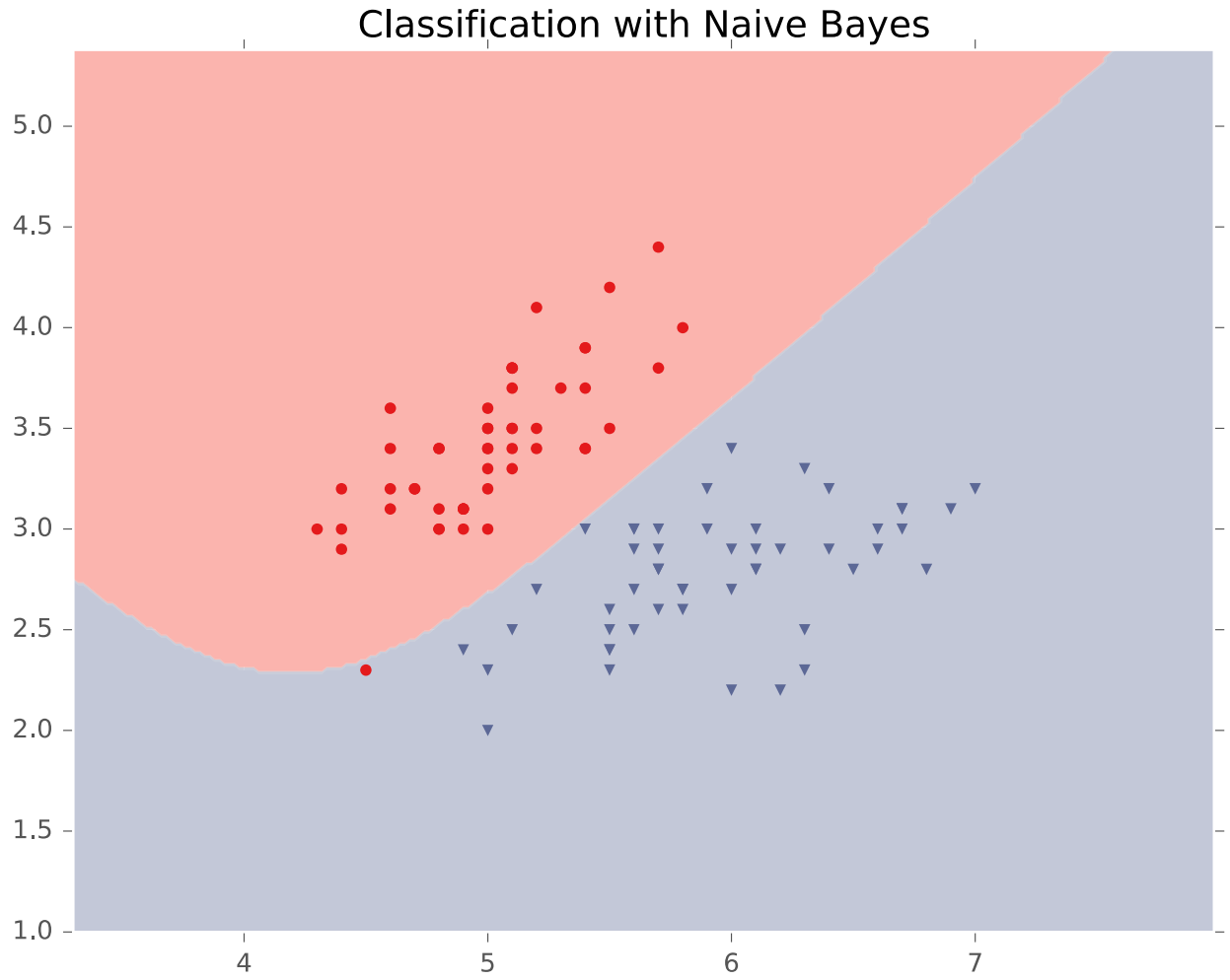


Figure courtesy of Matt Gormley

Bernoulli Naïve Bayes: Making Predictions

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

$$P(Y = 1|\boldsymbol{x}') \propto P(Y = 1)P(\boldsymbol{x}'|Y = 1)$$

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

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

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

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

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

$$P(Y = 1|\boldsymbol{x}') \propto P(Y = 1)P(\boldsymbol{x}'|Y = 1)$$

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

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

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

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 $\theta_{d,y} \sim \text{Beta}(\alpha, \beta)$
  - $\hat{\theta}_{d,y} = {}^{N_{Y=y,\, X_d=1}+(\alpha-1)}/_{N_{Y=y}+(\alpha-1)+(\beta-1)}$
    - $N_{Y=y}$ = # 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?

- ***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 much later when we discuss Bayesian networks

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