10-701: Introduction to Machine Learning Lecture 7 - Naïve Bayes

Henry Chai

9/20/23

Front Matter

- Announcements:
  - HW1 released 9/6, due 9/20 (today!) at 11:59 PM
  - HW2 released 9/20 (today!), due 10/4 at 11:59 PM
  - Instructor OH have been added to the course calendar
    - Monday's immediately after lecture
- Recommended Readings:
  - Murphy, Section 3.5

Bayesian Linear Regression

$p(y | x, w)$

$p(w)$

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \boxed{\epsilon \sim N(0, \sigma^2)} \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boxed{\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)}$$

then the distribution over $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim N(X\mathbf{0} + \mathbf{0} = \mathbf{0}, X\Sigma X^T + \sigma^2 I)$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\boldsymbol{0}, \Sigma)$$

then the *joint* distribution over $\boldsymbol{y}$ and $\boldsymbol{\omega}$ is

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\omega} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} X\Sigma X^T + \sigma^2 I & \Sigma X^T \\ X\Sigma & \Sigma \end{bmatrix}\right)$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\boldsymbol{0}, \Sigma)$$

then the *conditional* distribution of $\boldsymbol{\omega}$ given $\boldsymbol{y}$ is

$$\boldsymbol{\omega} \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{POST}, \Sigma_{POST})$$

where

$$\boldsymbol{\mu}_{POST} = \Sigma X^T (X \Sigma X^T + \sigma^2 I)^{-1} \boldsymbol{y},$$
$$\Sigma_{POST} = \Sigma - \Sigma X^T (X \Sigma X^T + \sigma^2 I)^{-1} X \Sigma$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the *conditional* distribution of $h(\boldsymbol{x}') = \boldsymbol{x}'^T \boldsymbol{\omega}$ given $\boldsymbol{y}$ is

$$h(\boldsymbol{x}') \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = \boldsymbol{x}'^T \Sigma X^T (X \Sigma X^T + \sigma^2 I)^{-1} \boldsymbol{y},$$

$$\Sigma_{PRED} = \boldsymbol{x}'^T \Sigma \boldsymbol{x}' - \boldsymbol{x}'^T \Sigma X^T (X \Sigma X^T + \sigma^2 I)^{-1} X \Sigma \boldsymbol{x}'$$

Kernelized Bayesian Linear Regression = Gaussian Process (GP)

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \cancel{\Sigma})$$

then the *conditional* distribution of $h(\boldsymbol{x}') = \boldsymbol{x}'^T \boldsymbol{\omega}$ given $\boldsymbol{y}$ is

$$h(\boldsymbol{x}') \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

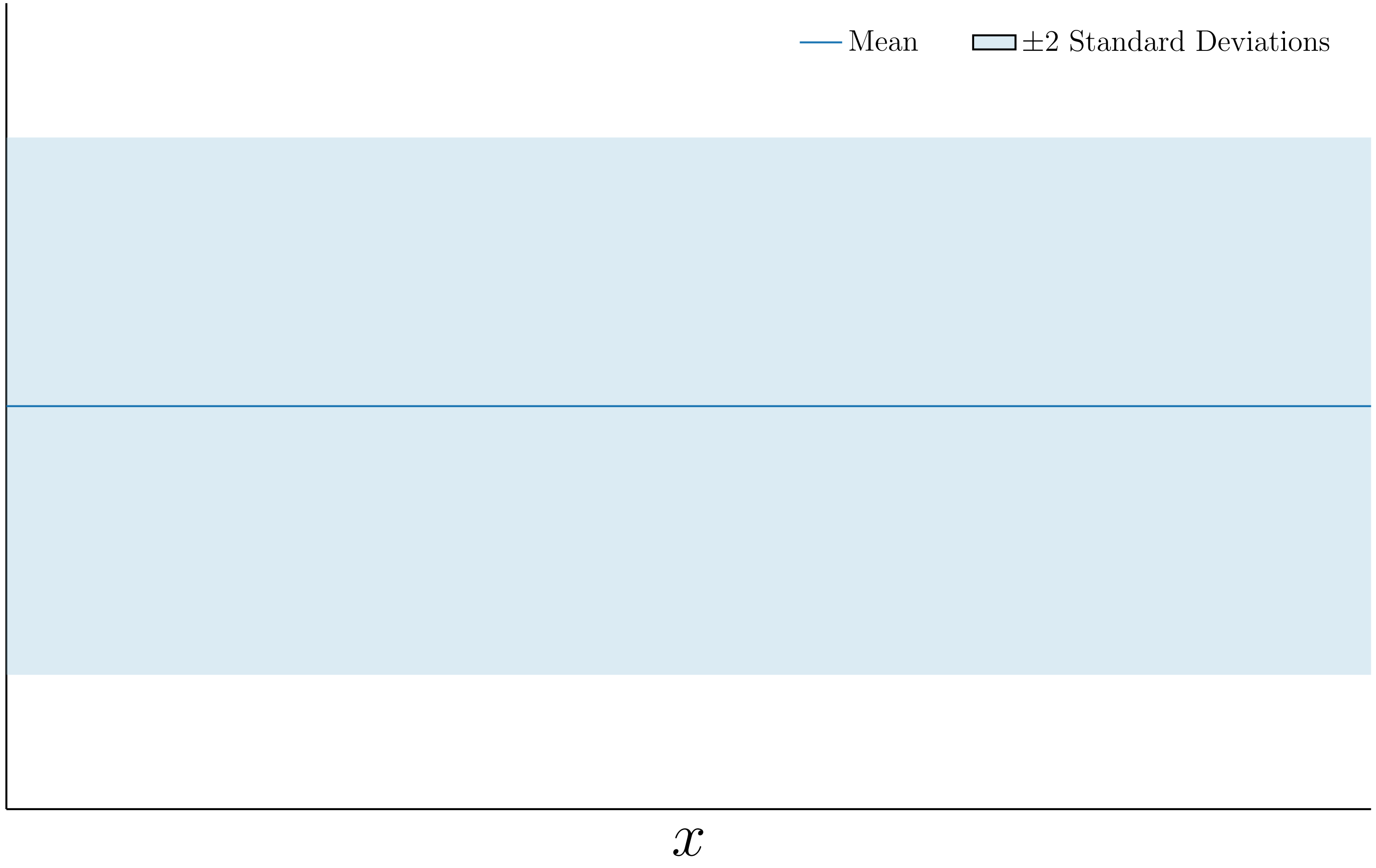
$$K(\boldsymbol{a}, \boldsymbol{b}) = \cancel{\Phi}(\boldsymbol{a})^T \Sigma \cancel{\Phi}(\boldsymbol{b})$$

$$\boldsymbol{\mu}_{PRED} = K(\boldsymbol{x}', X)(K(X,X) + \sigma^2 I)^{-1} \boldsymbol{y},$$

$$\Sigma_{PRED} = K(\boldsymbol{x}', \boldsymbol{x}') - K(\boldsymbol{x}', X)(K(X,X) + \sigma^2 I)^{-1} K(X, \boldsymbol{x}')$$

Gaussian Process (GP)

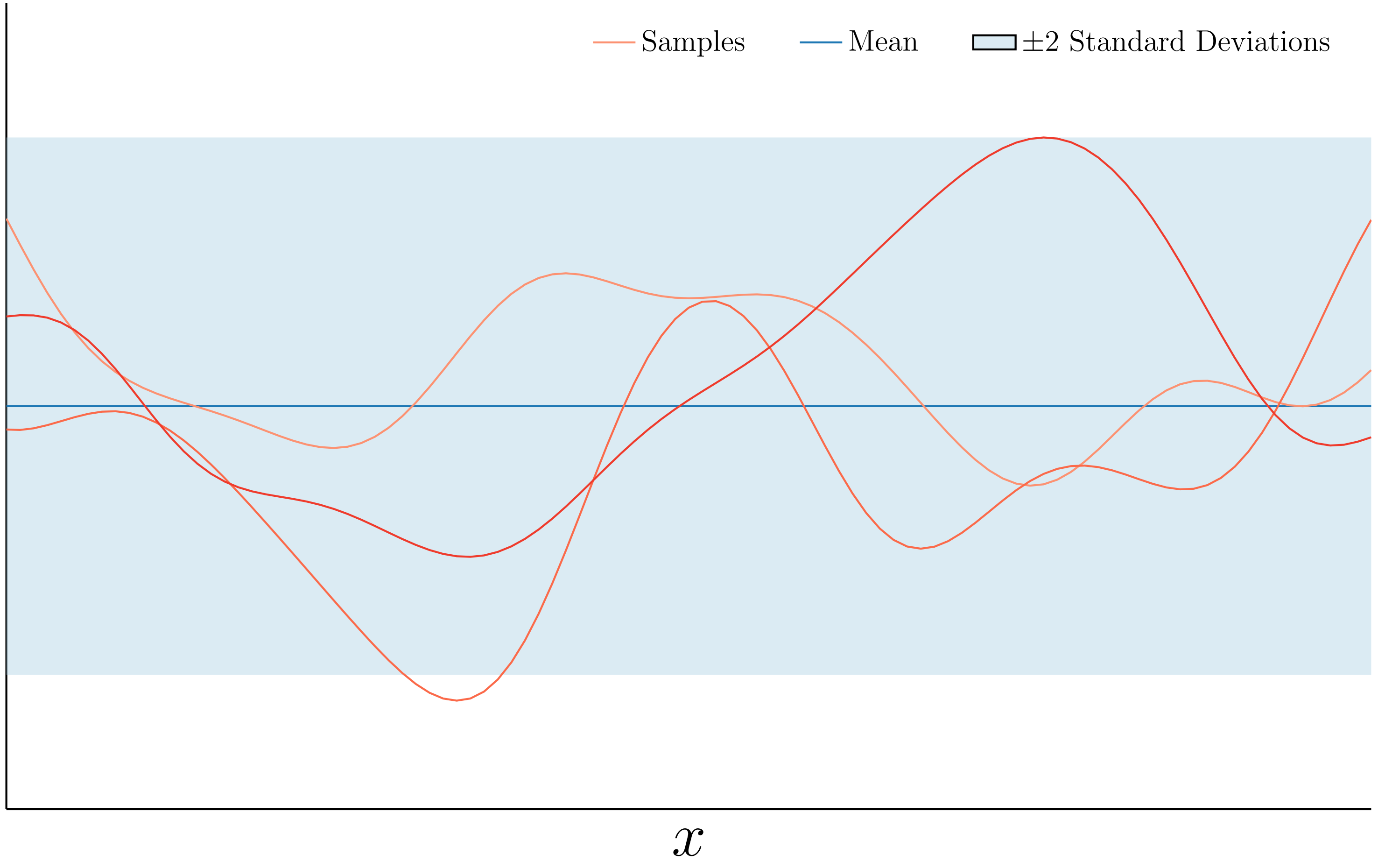
$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-(x - x')^2))$$



$$f \sim \mathcal{GP}(m, K) \rightarrow f(x) \sim \mathcal{N}\big(m(x), K(x, x)\big)$$

Gaussian Process (GP)

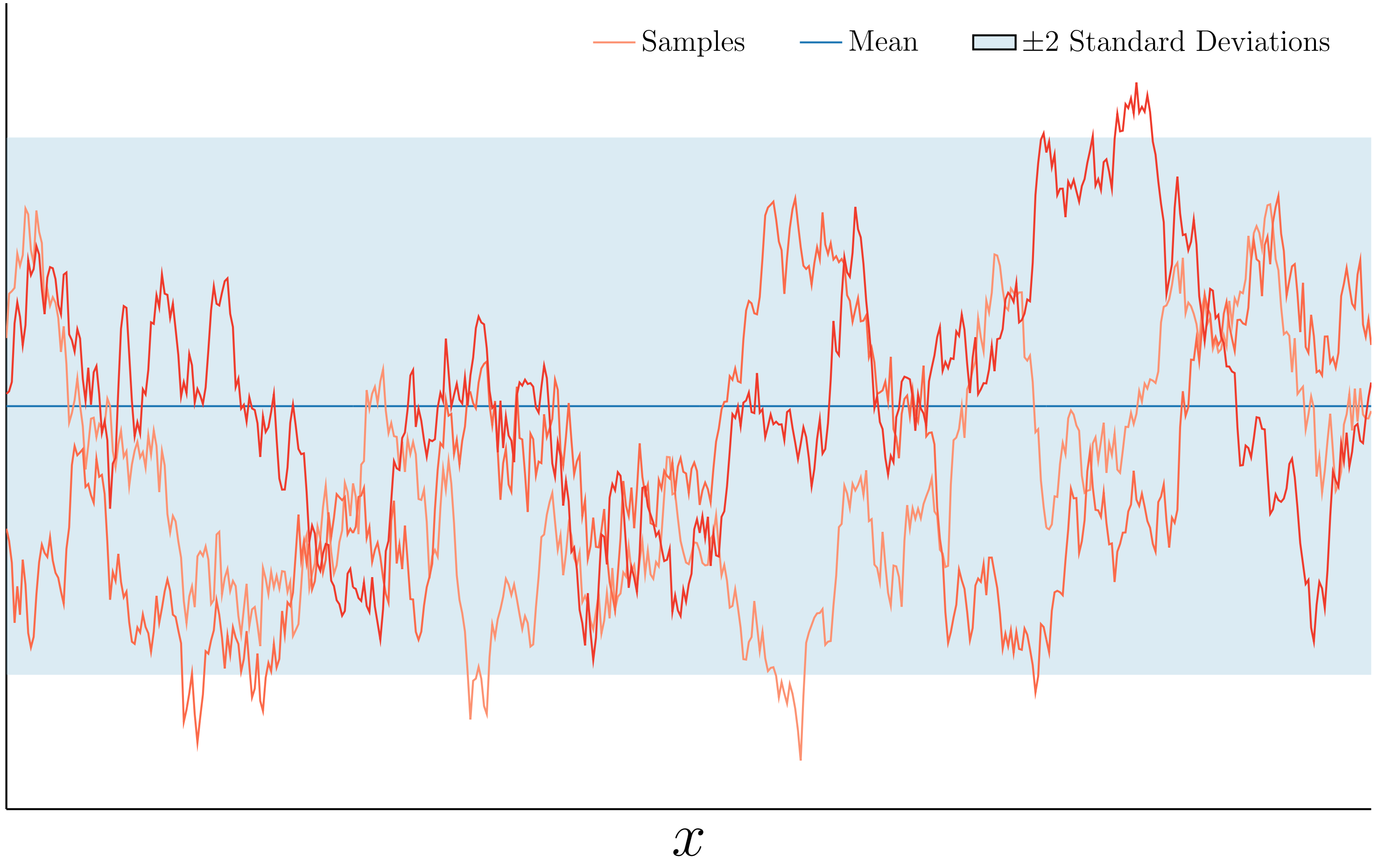
$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-(x - x')^2))$$



$$f \sim \mathcal{GP}(m, K) \rightarrow f(x) \sim \mathcal{N}\big(m(x), K(x, x)\big)$$

Gaussian Process (GP)

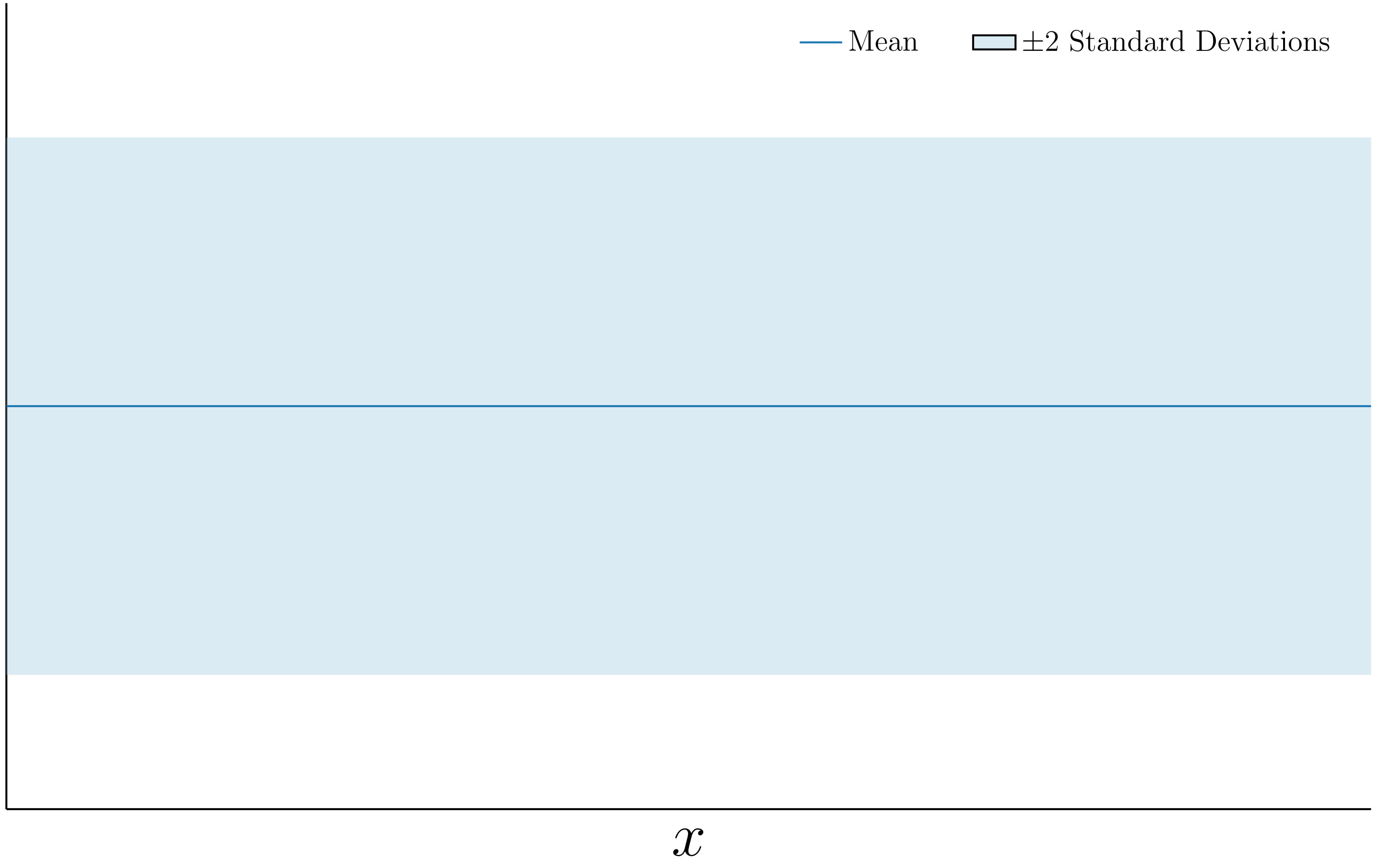
$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-\cancel{(}x - x'\cancel{)^2}))$$



$$f \sim \mathcal{GP}(m, K) \rightarrow f(x) \sim \mathcal{N}\big(m(x), K(x, x)\big)$$

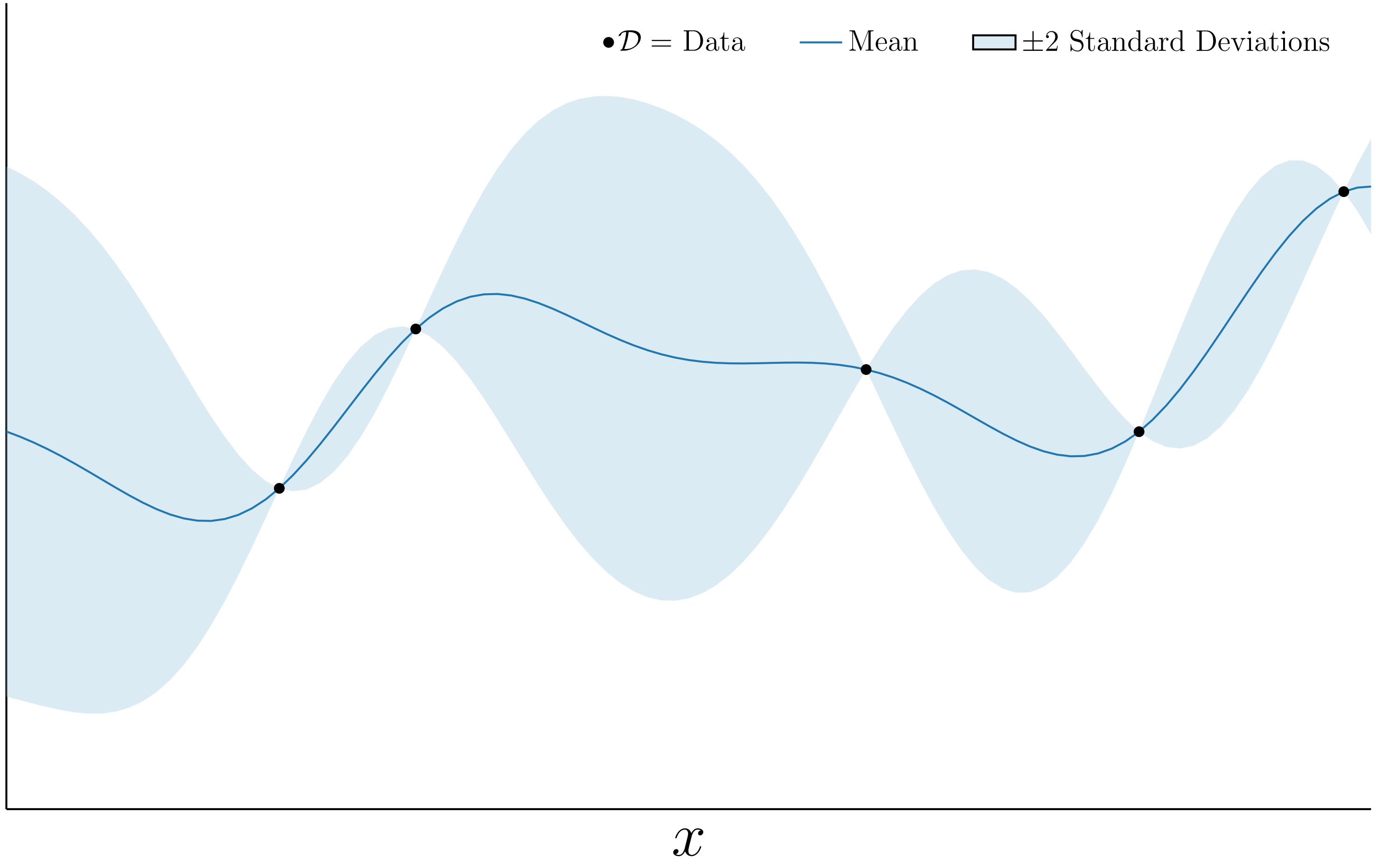
GP Prior

$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-(x - x')^2))$$



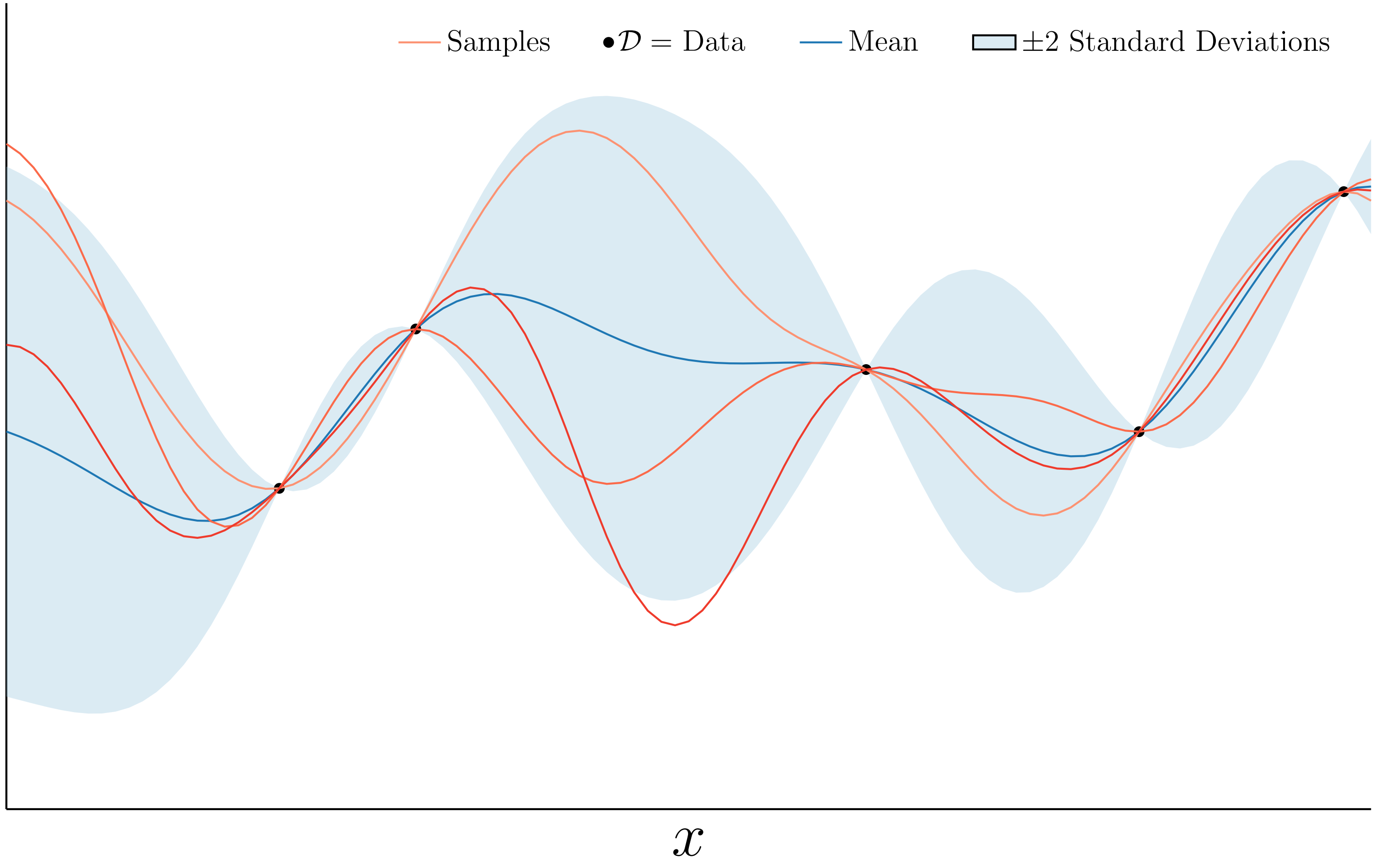
GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$$



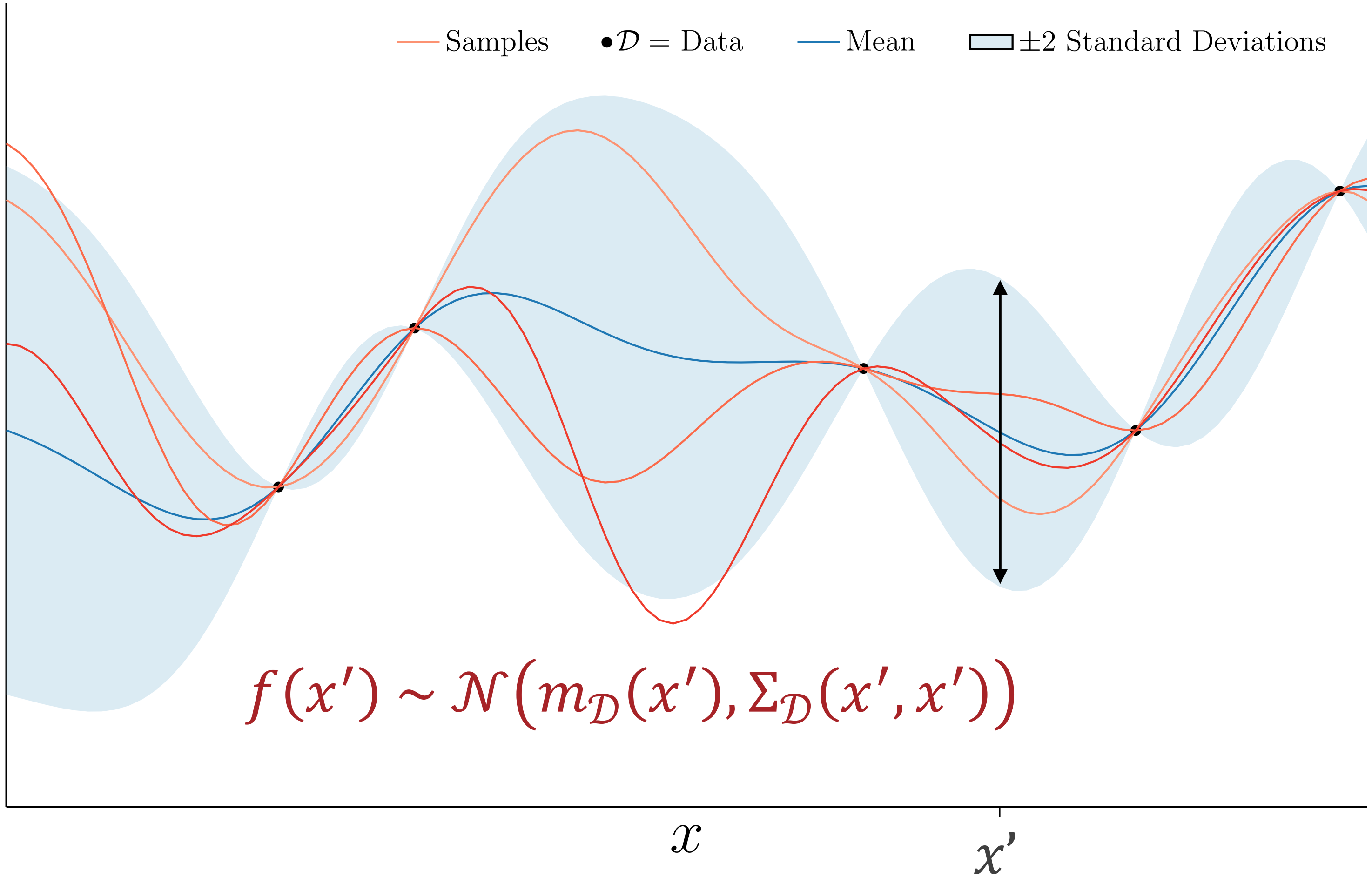
GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$$



GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$$



Text Data

- https://www.nytimes.com/2023/09/19/us/politics/senate-dress-code-fetterman-schumer.html
- https://americanwirenews.com/slobs-of-the-world-unite-schumer-changes-senate-dress-code-to-accommodate-fetterman/
- https://triblive.com/news/pennsylvania/u-s-senate-loosens-dress-code-scoring-win-for-casually-dressed-fetterman/
- https://www.theonion.com/fetterman-struggling-to-adapt-to-size-of-capitol-buildi-1849773669

Text Data

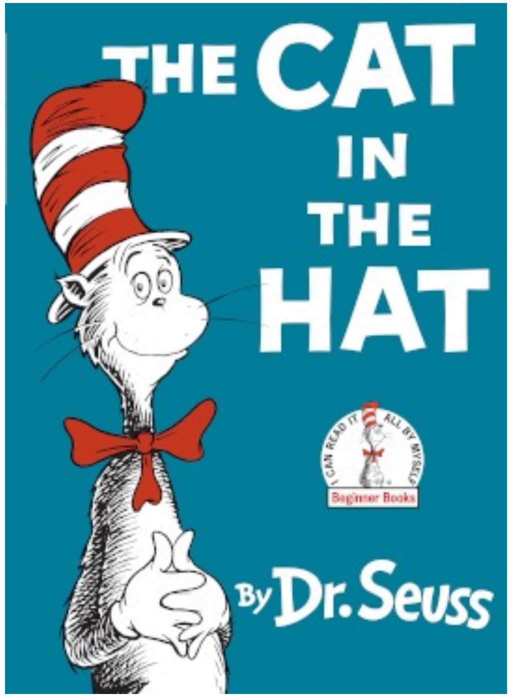
Bag-of-Words Model

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)

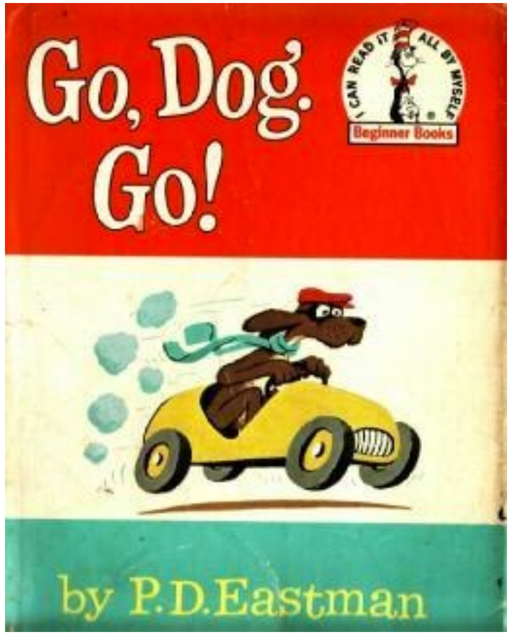


Source: https://en.wikipedia.org/wiki/The_Cat_in_the_Hat#/media/File:The_Cat_in_the_Hat.png

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)

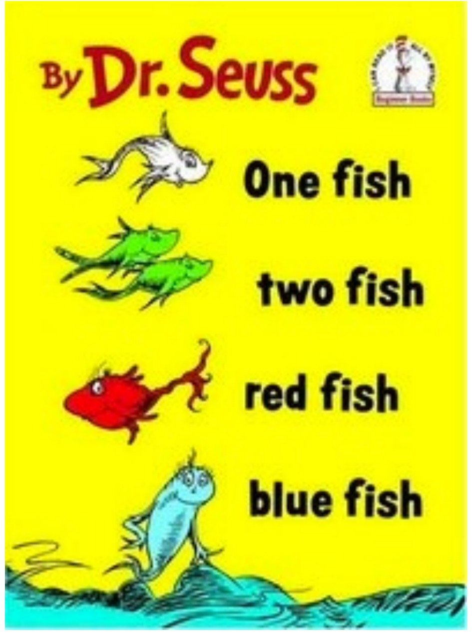


Source: https://en.wikipedia.org/wiki/Go,_Dog._Go!#/media/File:Go_Dog_Go.jpg

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)

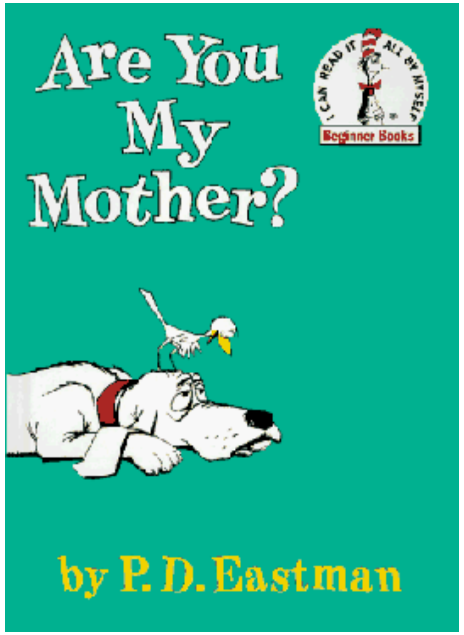


Source: https://en.wikipedia.org/wiki/One_Fish,_Two_Fish,_Red_Fish,_Blue_Fish#/media/File:One_Fish_Two_Fish_Red_Fish_Blue_Fish_(cover_art).jpg

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 (later)
  - Option 2 - Use Bayes' rule (today!):

likelihood

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

prior

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!):

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

$P(x_i = 1 \mid Y=1) \Rightarrow P(x_i = 0 \mid Y=1) = 1 - P(x_i=1 \mid Y=1)$

$P(x_i = 1 \mid Y=0)$

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")	$P(X\|Y=1)$
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:

$$P(X|Y) = P(x_1 \cap x_2 \dots x_D | Y) = \prod_{d=1}^{P} P(x_d | Y)$$

- Pros:

Simplifies inference/computation (linear dependence on D)

Combats overfitting (training)

- Cons:

It's a frequently incorrect assumption

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
  - Assume iid data points
  - Parameters: $\pi = P(Y=1)$, $\theta_{d,y} = P(X_d = 1 \mid Y=y)$
- Write down an objective function
  - Maximize the log-likelihood
- Optimize the objective w.r.t. the model parameters
  - Solve for them in closed-form

Setting the Parameters via MLE

$$\ell_D(\pi,\theta) = \log\, p\left(D=\{(x^{(i)},y^{(i)})\}_{i=1}^{N} \mid \pi,\theta\right)$$

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

$$= \sum_{i=1}^{N} \log\left(p(x^{(i)} \mid y^{(i)},\theta)\, p(y^{(i)} \mid \pi)\right)$$

$$= \sum_{i=1}^{N} \log\left(p(y^{(i)} \mid \pi) \prod_{d=1}^{D} p(x_d^{(i)} \mid y^{(i)}, \theta_{d,1}, \theta_{d,0})\right)$$

$$= \sum_{i=1}^{N}\left(\log\left(\underbrace{p(y^{(i)} \mid \pi)}\right) + \sum_{d=1}^{D} \log\left(p(x_d^{(i)} \mid y^{(i)}, \theta_{d,1}, \theta_{d,0})\right)\right)$$

$$= \sum_{i=1}^{N}\left(\log\left(\pi^{y^{(i)}}(1-\pi)^{1-y^{(i)}}\right)\right)$$

$$+ \sum_{i:y^{(i)}=1} \sum_{d=1}^{D} \log\left(\theta_{d,1}^{x_d^{(i)}}(1-\theta_{d,1})^{1-x_d^{(i)}}\right) + \sum_{i:y^{(i)}=0} \cdots$$

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

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)

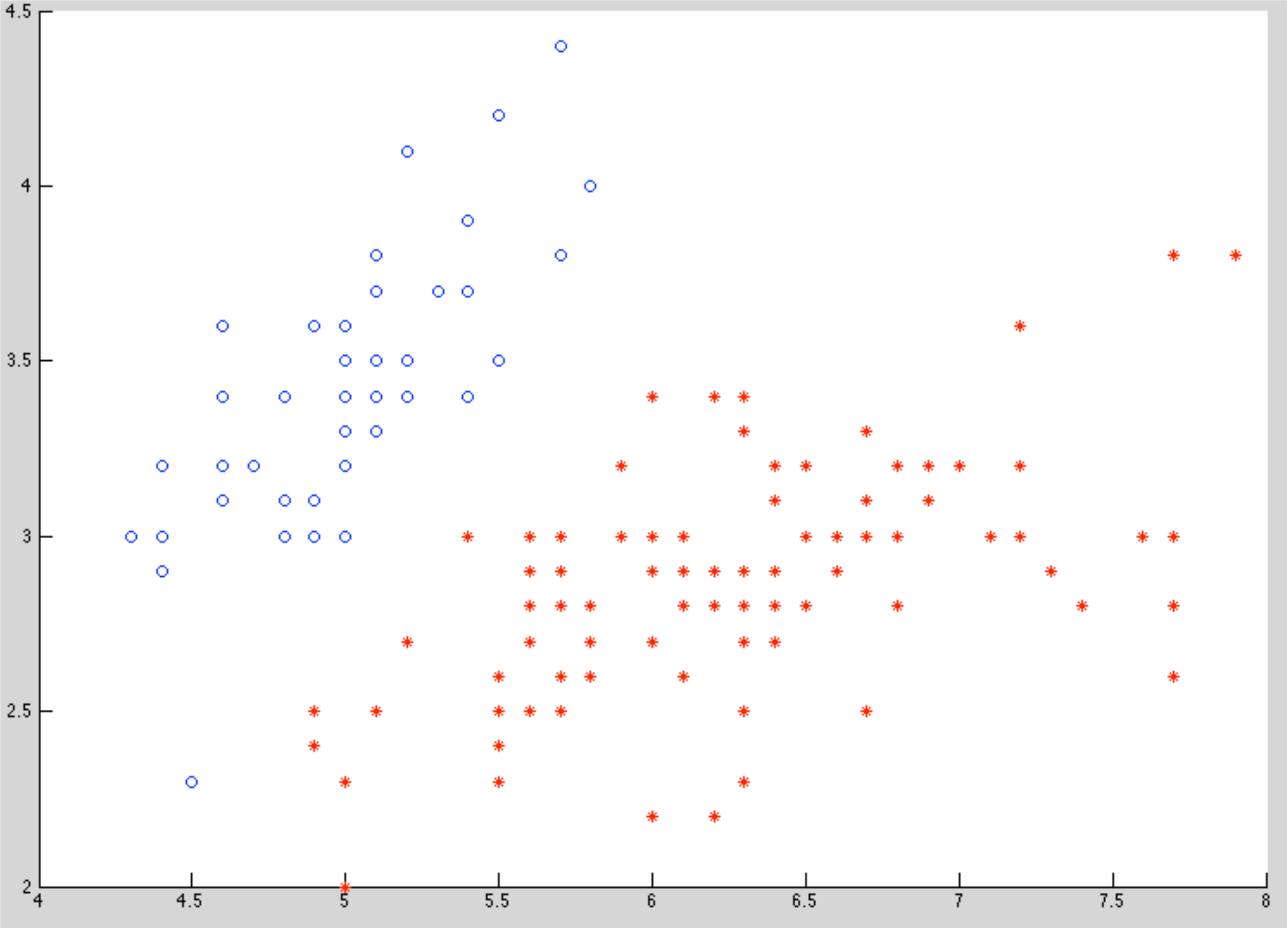


Figure courtesy of William Cohen

Visualizing Gaussian Naïve Bayes (2 classes)

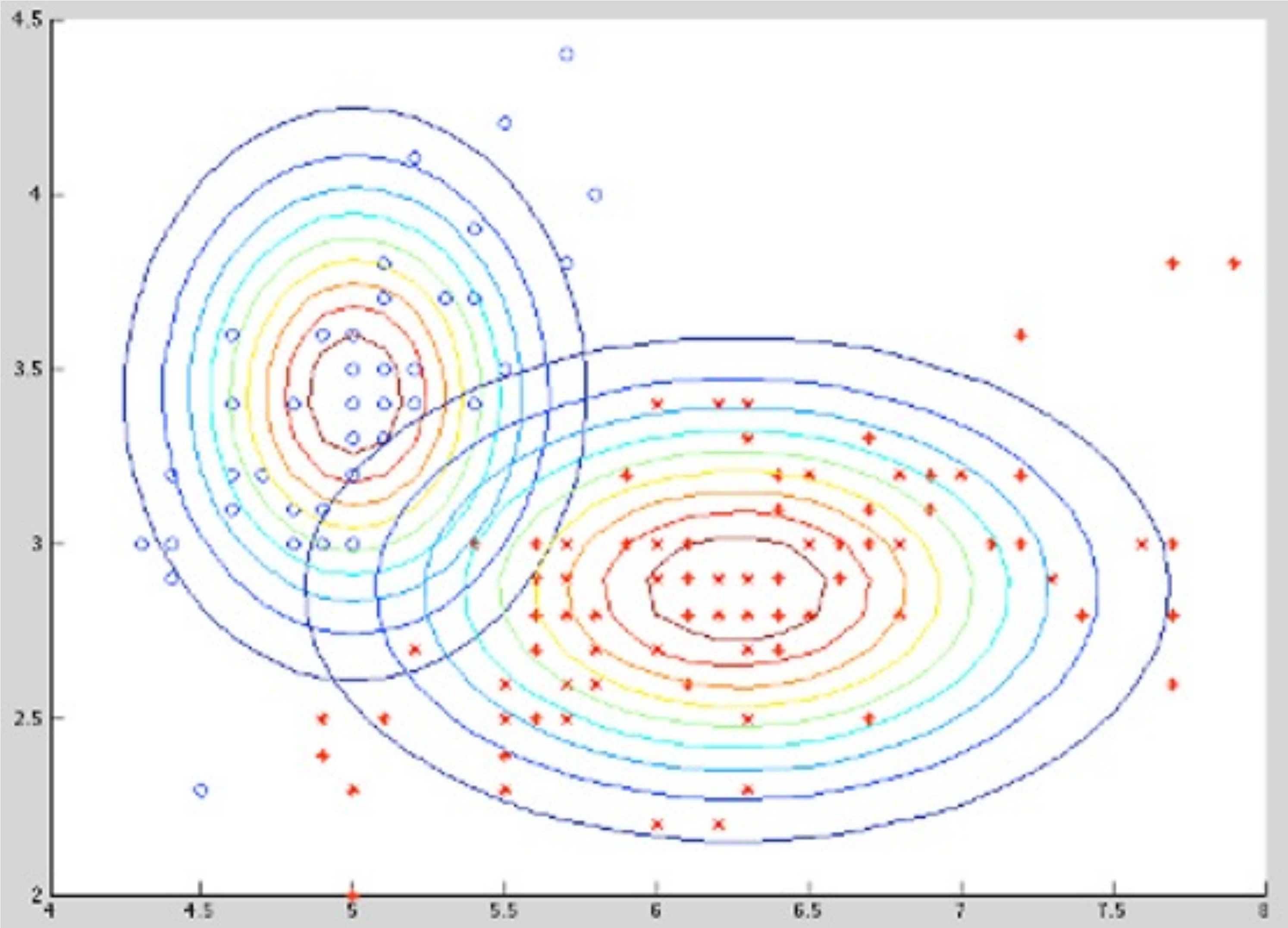


Figure courtesy of William Cohen

$\rightarrow \hat{\sigma}^2_{d,1} = \hat{\sigma}^2_{d,0}$

$\rightarrow \hat{\sigma}^2_{x_1,1} = \hat{\sigma}^2_{x_2,1}$

Visualizing Gaussian Naïve Bayes (2 classes, equal variances)

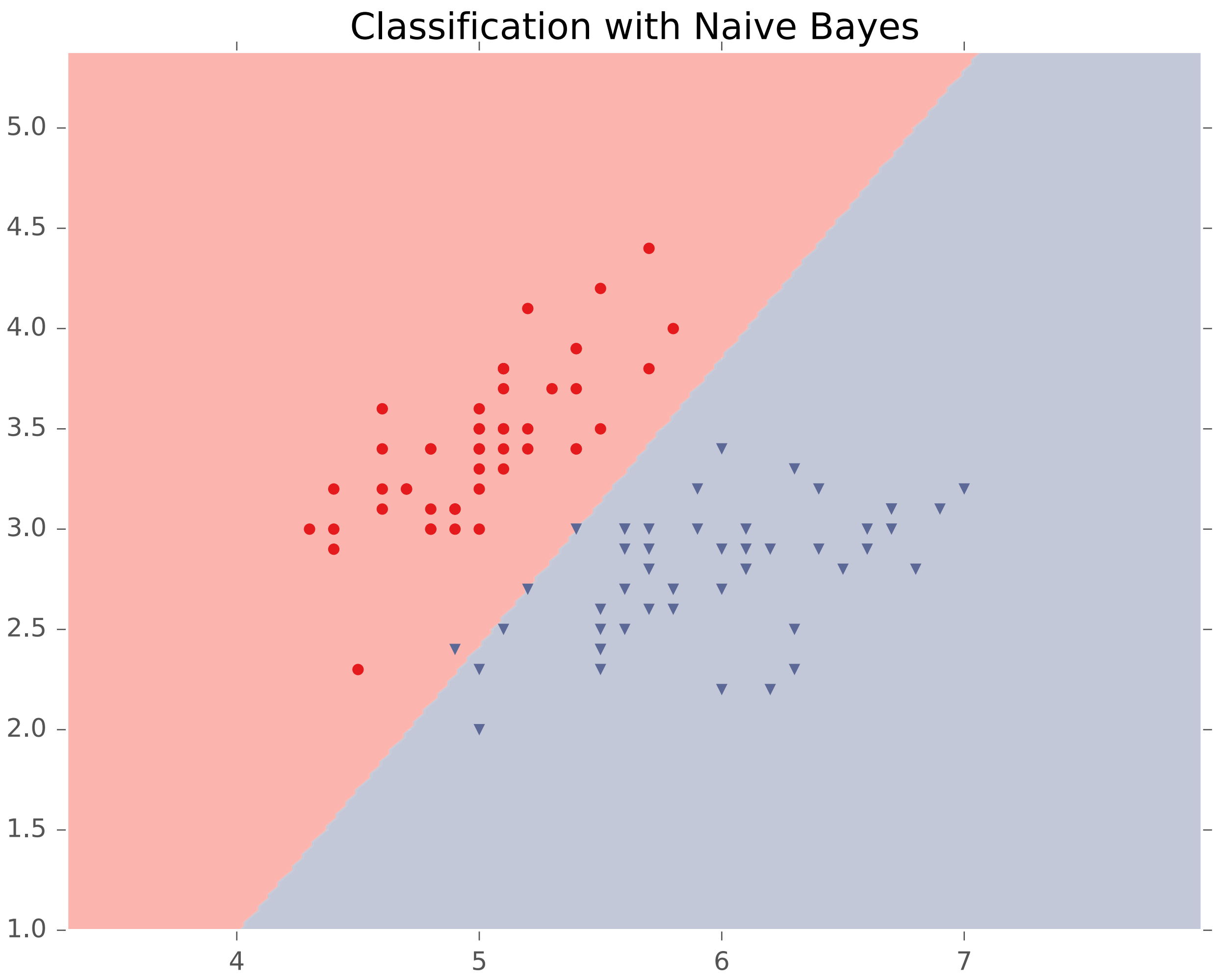


Figure courtesy of Matt Gormley

Visualizing Gaussian Naïve Bayes (2 classes, learned variances)

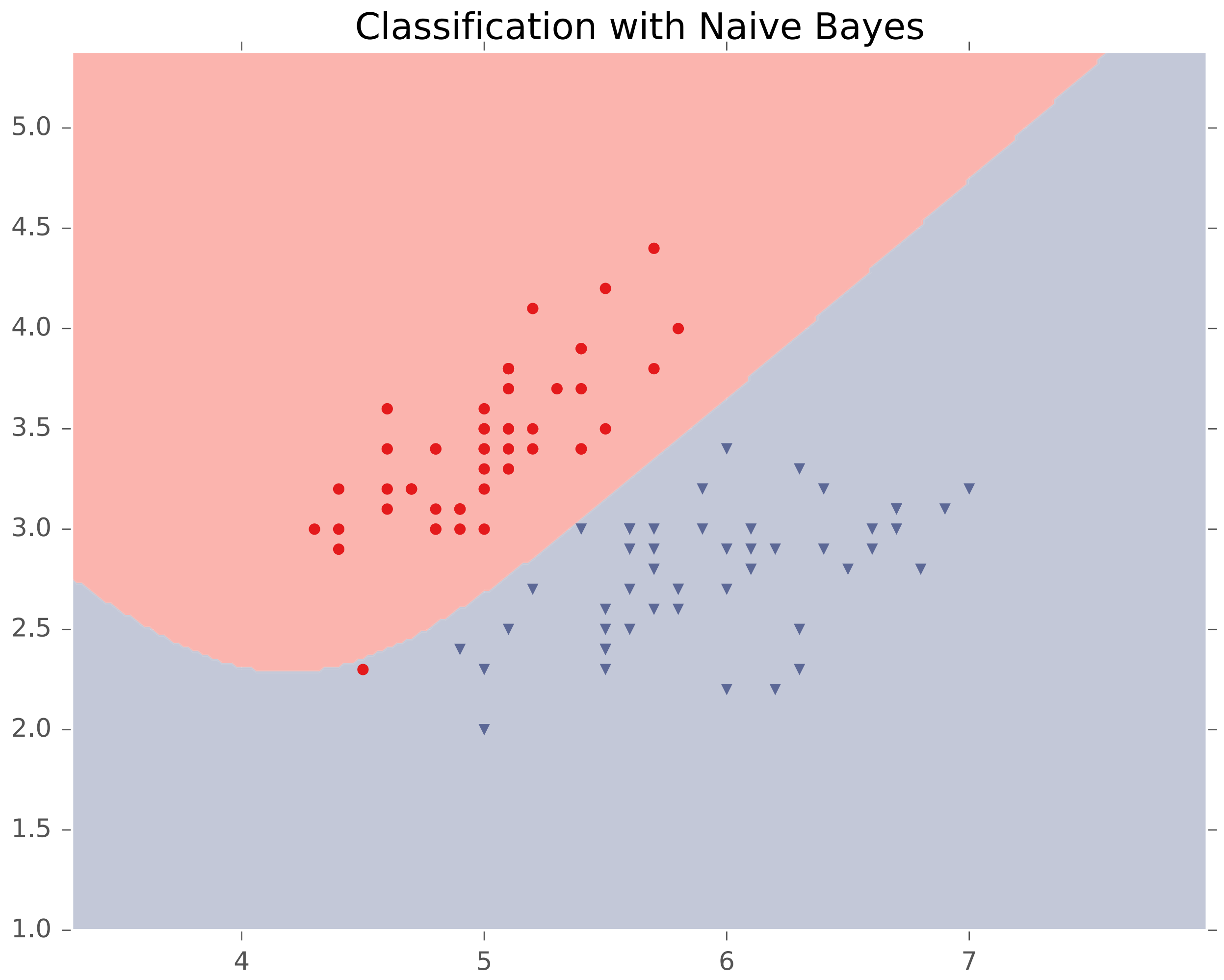


Figure courtesy of Matt Gormley

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

Bernoulli Naïve Bayes: Making Predictions

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

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

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

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

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

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 next week 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