

Bayesian Classification and Regression

DS6040 Fall 2024
Teague R. Henry

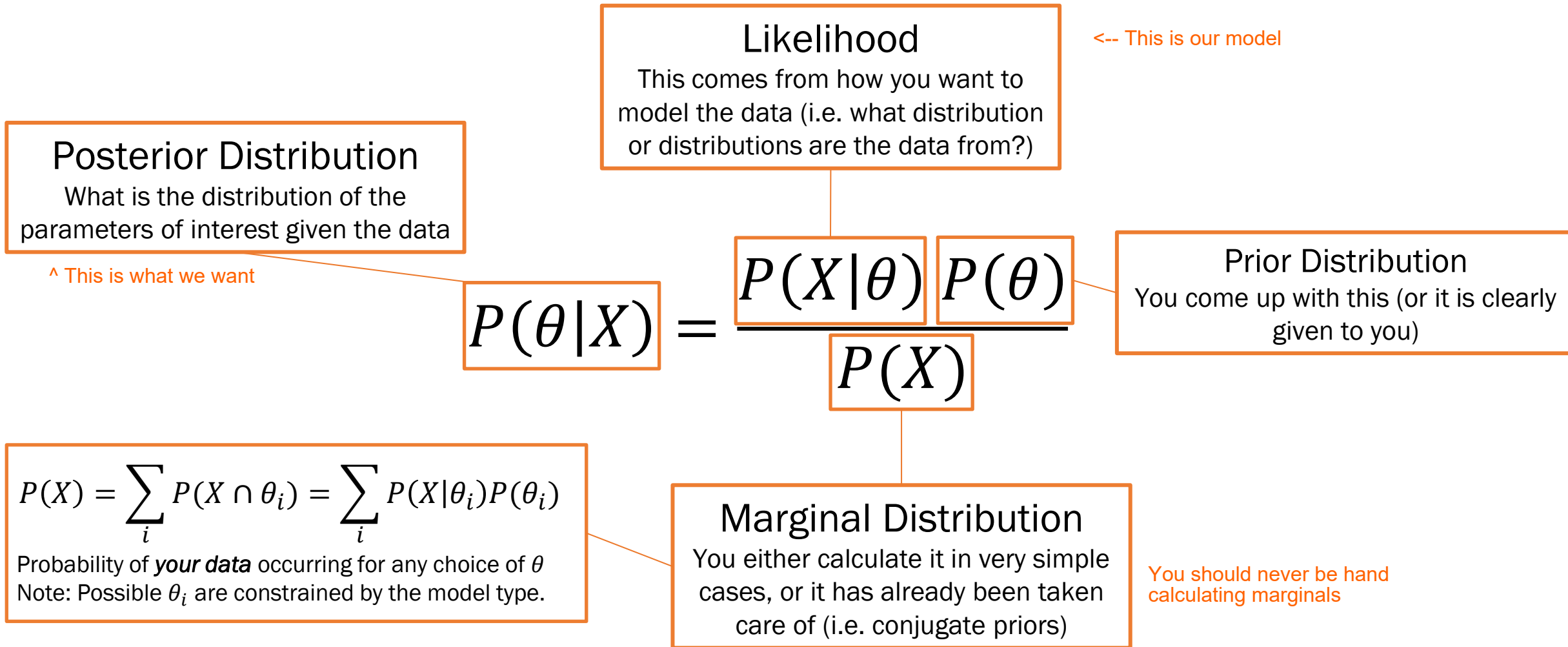


SCHOOL *of* DATA SCIENCE

Outline


- Prior Review
- Evaluating Posterior Distributions
- Linear Classification
 - Linear Discriminant Analysis
 - Quadratic Discriminant Analysis
- Linear Regression

Where Do You Get These Pieces?





Prior Review

Conjugate Priors

- A prior is conjugate if the posterior is of the same distribution type as the prior
- If likelihood $P(X|\theta)$ is $N(\mu, \sigma^2)$ with known σ^2 , then conjugate prior for μ is normal
 - Which means that the posterior distribution of μ is normal
- If likelihood is binomial, the conjugate prior for p is a beta distribution
 - Which means that the posterior distribution of p is a beta.
- Priors are specific to parameters
 - If a likelihood has multiple parameters, then a prior for each of the parameters need to be specified (or a multivariate prior needs to be specified).
 - Incorrect statement: The conjugate prior for a normal likelihood is normal 
 - Correct statement: The conjugate prior for the μ parameter of a normal likelihood with known variance is normal.
- **Conjugate priors mean you can skip using Bayes Theorem directly to the posterior!**

Prior Review

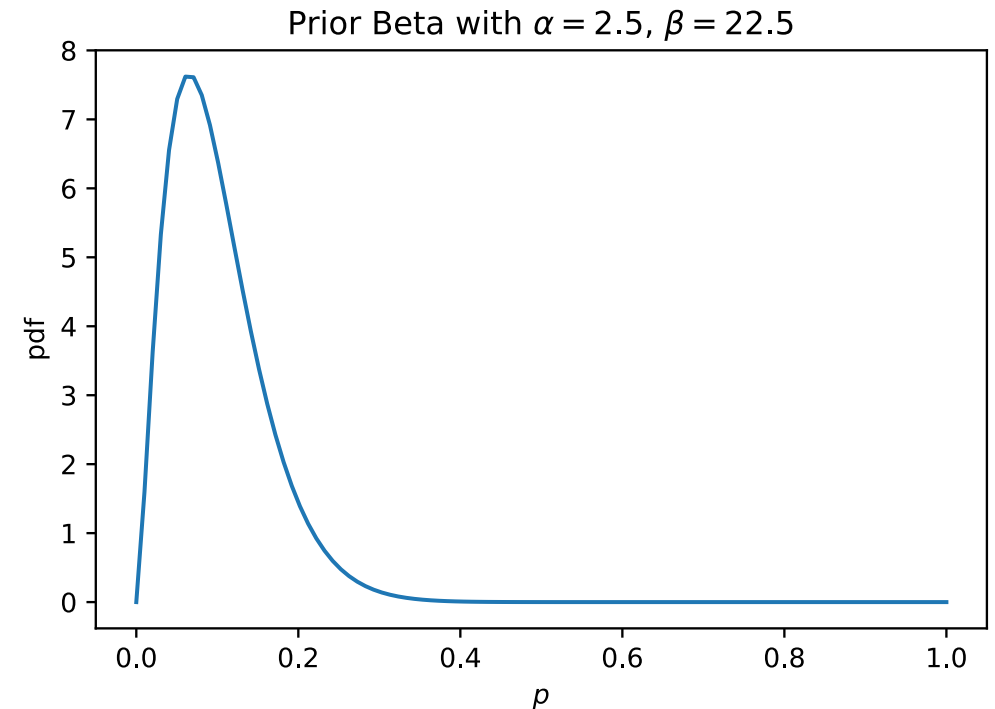
Non-Informative Priors

- A non-informative prior is a prior that doesn't carry much subjective information
- “Flat Priors”
 - Constant priors $p(\theta) = c$, for any constant c (These disappear from Bayes Theorem)
 - Extremely flat distributions (i.e. $p(\mu) \sim N(0, 10000)$) 
- Jeffreys Priors
 - Non-informative priors that are invariant under reparameterizations
 - Fixes the issue of non-informative priors being highly informative when you change the model to a statistically equivalent, but reparametrized, version
- Issues with Non-informative Priors
 - They still are highly informative at low sample sizes 
 - Multivariate non-informative priors have increasing issues as the number of variables increase

Evaluating Posterior Distributions

You are modelling the failure rate of a new type of processor for your company.

- Out of the 10 test processors, 3 have failed so far. (Binomial Likelihood!)
- Previous generations of similar processors had an average failure rate of .1, with a variance of 0.003461538.
 - This corresponds to a Beta prior with $\alpha = 2.5, \beta = 22.5$ 💬



Evaluating Posterior Distributions

Using the Beta conjugate prior for a Binomial likelihood leads to a posterior $Beta(5.5, 29.5)$

$$P(p|k, n) = Beta(\alpha + k, n - k + \beta)$$

How can we characterize the posterior?

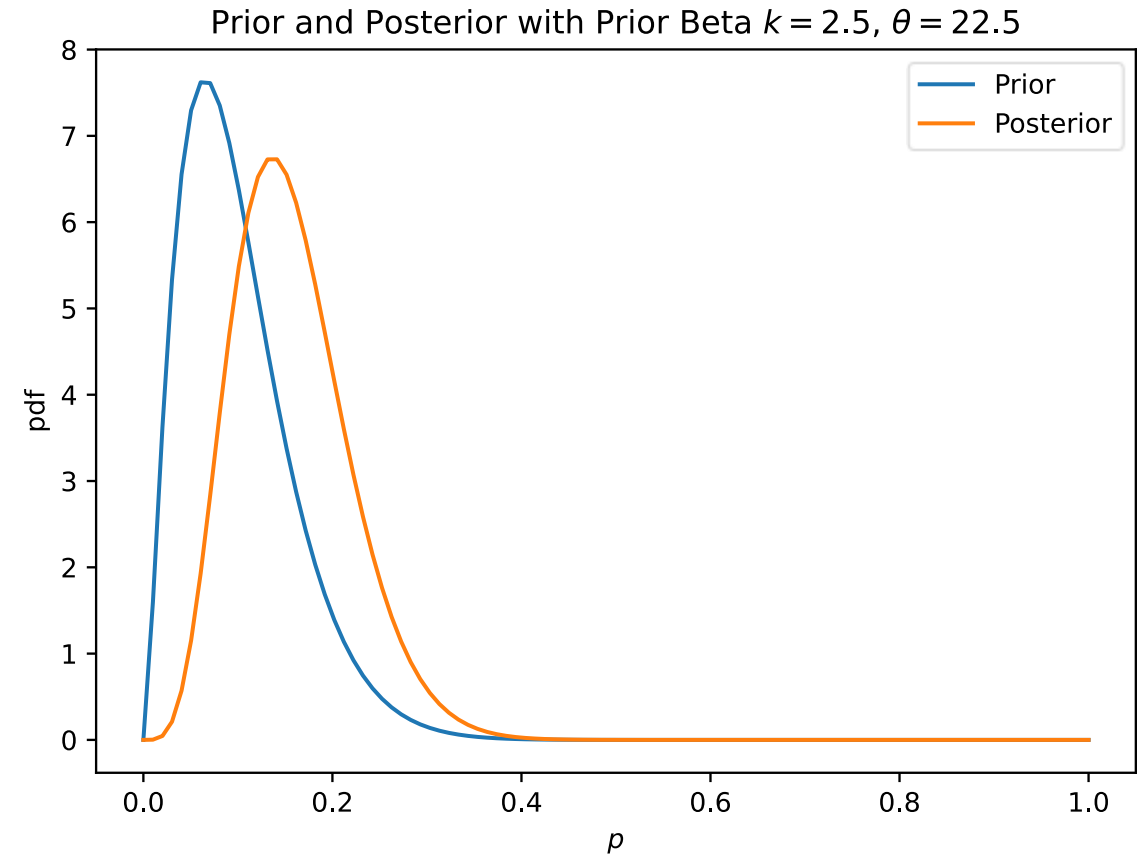
We need to do this in terms of numbers. How? Expected value potentially

- EAP (Expected a posteriori value)
 - Expected value of the posterior...
- MAP (Modal a posteriori value)
 - Mode of the posterior

This is the most common value of the posterior

- Credible Intervals

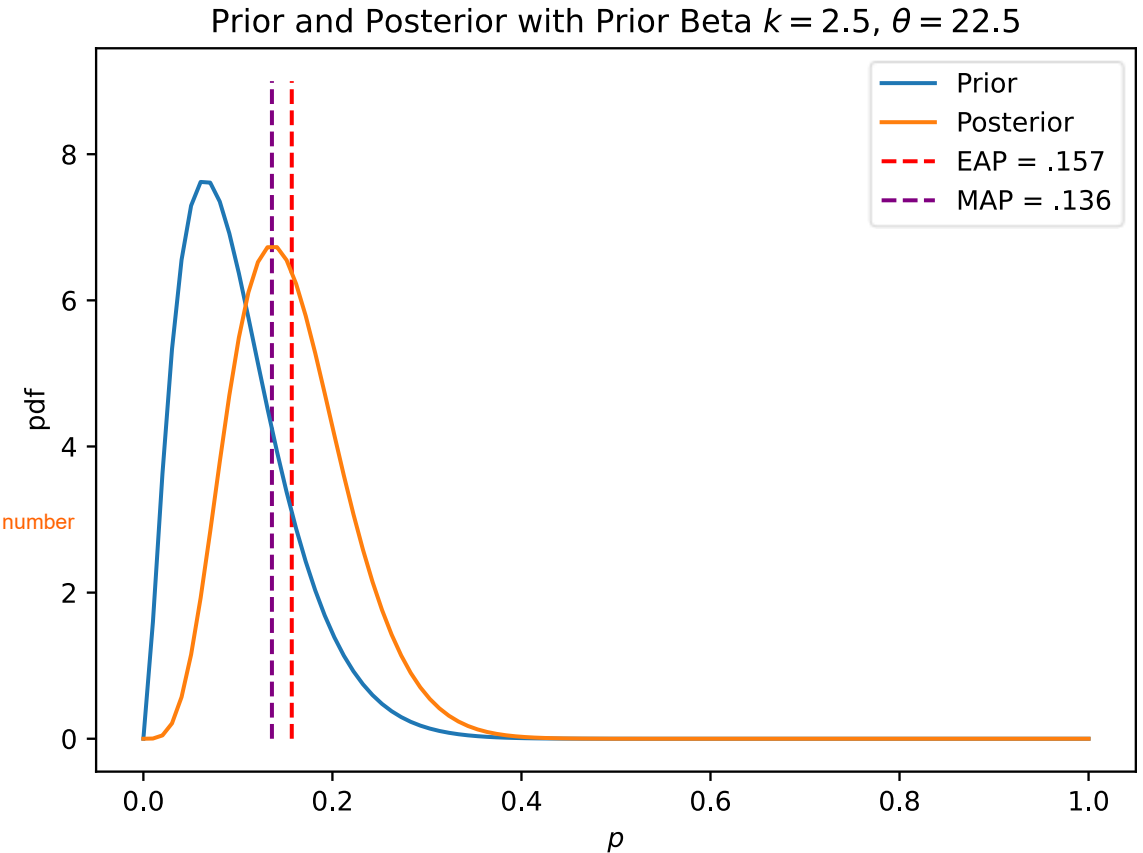
These are the errors of the probability that contain a portion of the map



Evaluating Posterior Distributions

Posterior $\sim \text{Beta}(5.5, 29.5)$

- EAP (Expected a posteriori value)
 - EAP = .157 5.5/29.5
 - On average across multiple generations of chips like this, we should see a failure rate of .157 b/c this is the expected value: If we were to do this an infinite number of times, what is the expected rate.
- MAP (Modal a posteriori value)
 - MAP = .136
 - Our best guess for the failure rate of the current generation of chips is .136



Evaluating Posterior Distributions

Posterior $\sim \text{Beta}(5.5, 29.5)$

Credible Interval

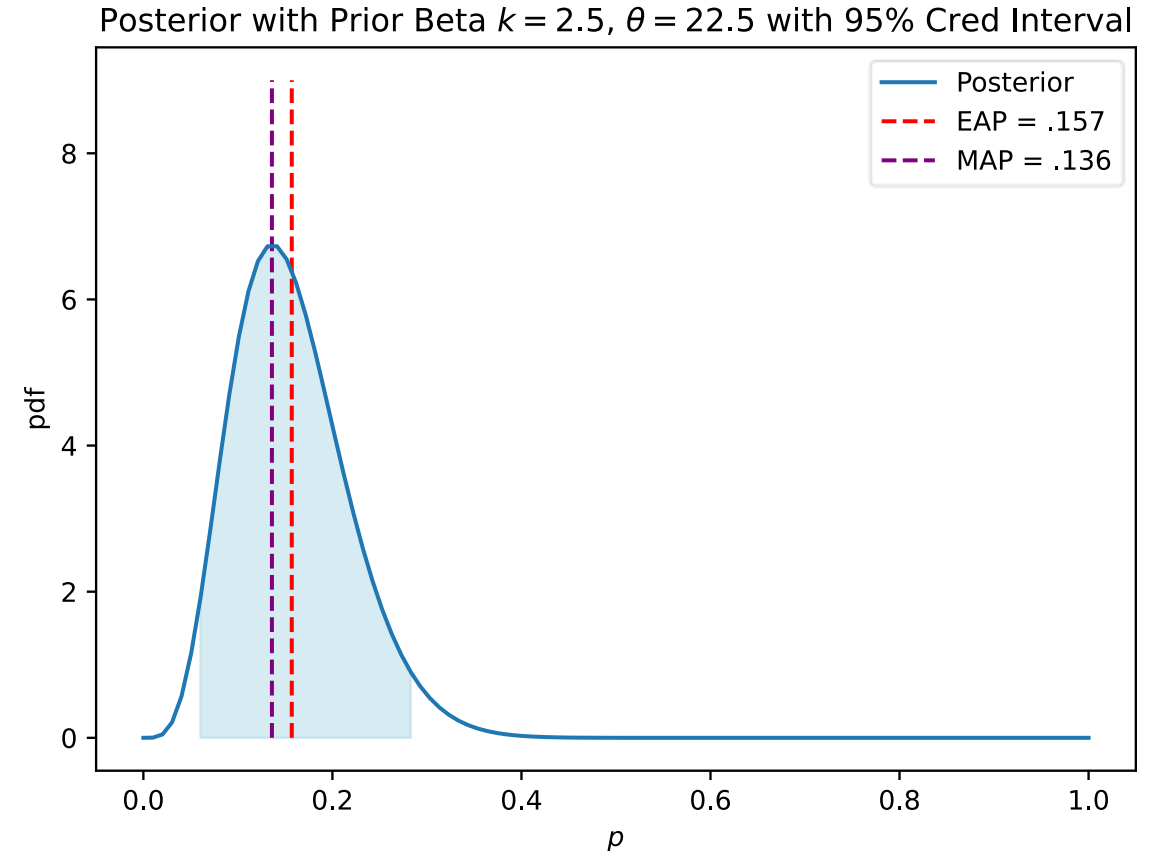
- Equal Tailed Credible Interval
 - From .025 quantile to .975 quantile
- 95% CredI = [.06, .29]

We have a 95% probability that our actual failure rate will fall in-between these two values.

- **Not unique**

- Equal Tailed Credible Intervals are one option of several

This is b/c you can slide the shaded area around and still capture a 95% credible Interval. Credible intervals are what confidence intervals want to be. They let you say things like we are "pretty sure"



Bayesian Classification

Supervised Classification –

- Given data that has discrete class labels, how can we best label new observations?

Unsupervised Classification –

- Given data without discrete class labels, can we cluster them into meaningful groups?

Today we are focusing on supervised classification. b/c unsupervised classification is a different ball game and a different field.

Bayes Optimal Classifiers

Consider the problem of classifying cells as cancerous/non-cancerous:

- You are given data (X) on cancerous/non-cancerous ($C = 1$ or 0) cells
- You want to build a classifier that can take new observations and predict if those new cells are cancerous/non-cancerous.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

What do you need here?

There are two things that we need: 1) we need to know the prior and 2) we need to know the likelihood

- What are the prior probabilities that a cell is cancerous/non-cancerous $P(C)$
- How is your data (X) distributed if the cell is cancerous/non-cancerous $P(X|C)$
- *From the likelihood and the prior, we can calculate the marginal $P(X)$*
 - *What is the probability of your observed data, regardless of cancer status*

Bayes Optimal Classifiers

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

You observe X_{new} . What would be the class assignment that minimizes the chance of misclassification?

This is not a trick question: The best guess for the class assignment of X_{new} is the choice of C that maximizes $P(C|X)$

Bayes Optimal Classifiers

We want a classification rule that can give us the probability of group membership given our data...

Use of Bayes is appropriate here, but...

In order to produce a real classifier, we need to specify what we mean by $P(X|C)$.

In other words, we need to know what the distribution of our data is for every group we are interested in...

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

We also need to specify priors for our group membership

This will be fairly easy to calculate when we are working with classifiers

Quadratic/Linear Discriminant Analysis

Let's assume that each class in our data comes from a multivariate normal distribution.

$$P(X|C = i) \sim N(\mu_i, \Sigma_i) \text{ Likelihood!}$$

Where μ_i is the vector of means for class i , and Σ_i is the covariance matrix for class i .

This setup is known as a quadratic discriminant analysis

If $\Sigma_i = \Sigma$ for all i , then this is a linear discriminant analysis.

Note: Inherent in this setup is the assumption that we know the values of μ_i and Σ_i

Quadratic/Linear Discriminant Analysis

$$P(X|C = i) \sim N(\mu_i, \Sigma_i) \quad \text{Likelihood!}$$

What do we need?

- Priors on $C = i$
- Calculate out our marginal distribution $P(X)$

Say we have 3 classes with $P(C = i) = .2, .3, .5$ respectively...

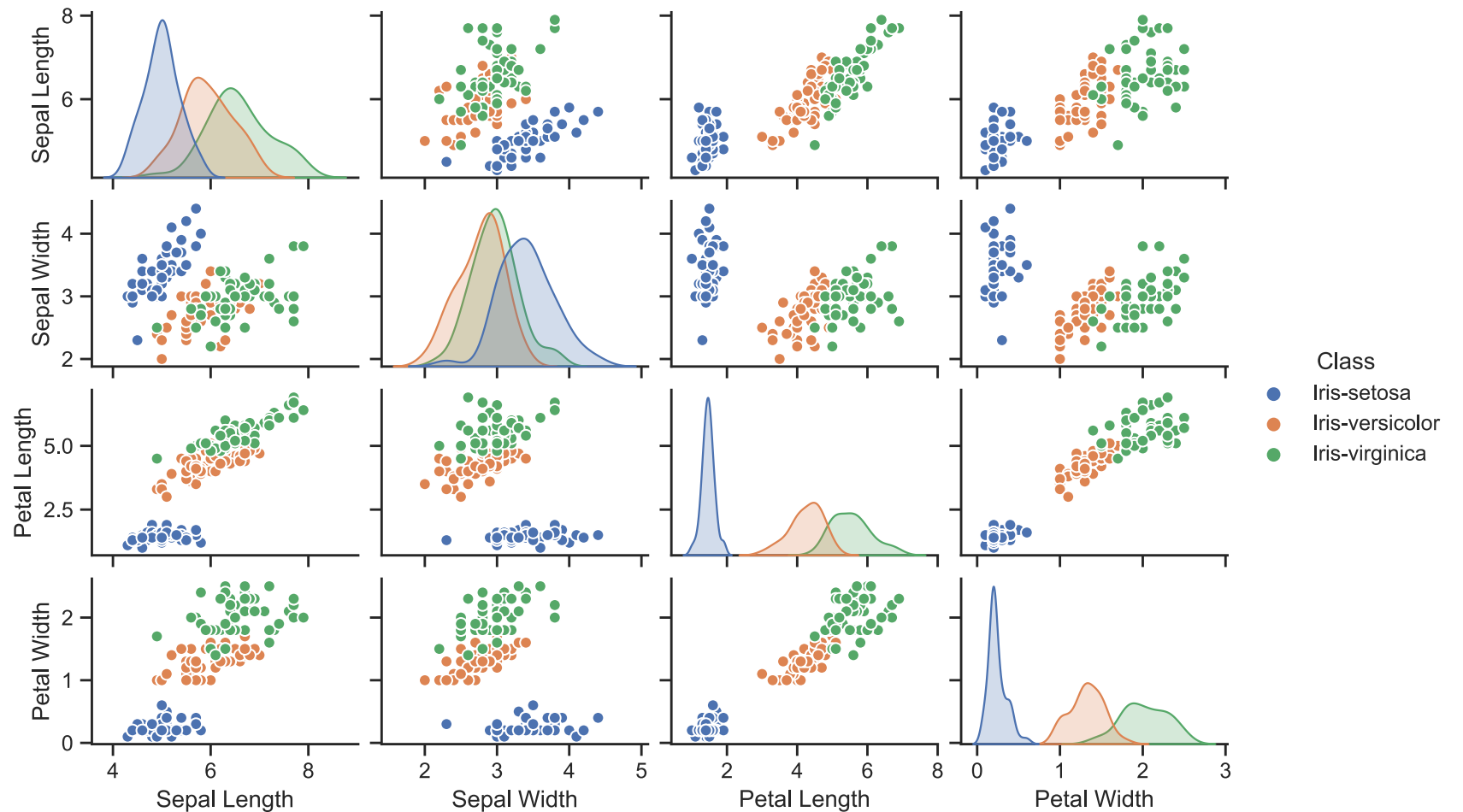
$$P(C = 1|X) = \frac{\text{Normal PDF}(X, \mu_1, \Sigma_1) * .2}{.2 * \text{Normal PDF}(X, \mu_1, \Sigma_1) + .3 * \text{Normal PDF}(X, \mu_2, \Sigma_2) + .5 * \text{Normal PDF}(X, \mu_3, \Sigma_3)}$$

Classifying Irises

Fisher's Irises

- 3 Classes
- 4 Variables
 - Sepal Length
 - Sepal Width
 - Petal Length
 - Petal Width

Let's build some classifiers for these irises...



LDA - Irises

LDA specifies that each class of irises have unique vectors of means, but the same covariance matrix.

μ for each class

Class	Sepal Length	Sepal Width	Petal Length	Petal Width
Iris-setosa	5.006	3.418	1.464	0.244
Iris-versicolor	5.936	2.77	4.26	1.326
Iris-virginica	6.588	2.974	5.552	2.026

Σ overall

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	0.686	-0.039	1.274	0.517
Sepal Width	-0.039	0.188	-0.322	-0.118
Petal Length	1.274	-0.322	3.113	1.296
Petal Width	0.517	-0.118	1.296	0.582

LDA - Irises

Now, we have our μ_i and Σ . That is all we need from the current dataset...

For a new observation Y

$$P(C = i|Y) = \frac{P(Y|\mu_i, \Sigma)P(C = i)}{\sum_{k \in \{1,2,3\}} P(Y|\mu_k, \Sigma)P(C = k)}$$

The Bayes Optimal Decision rule is to assign the highest probability class

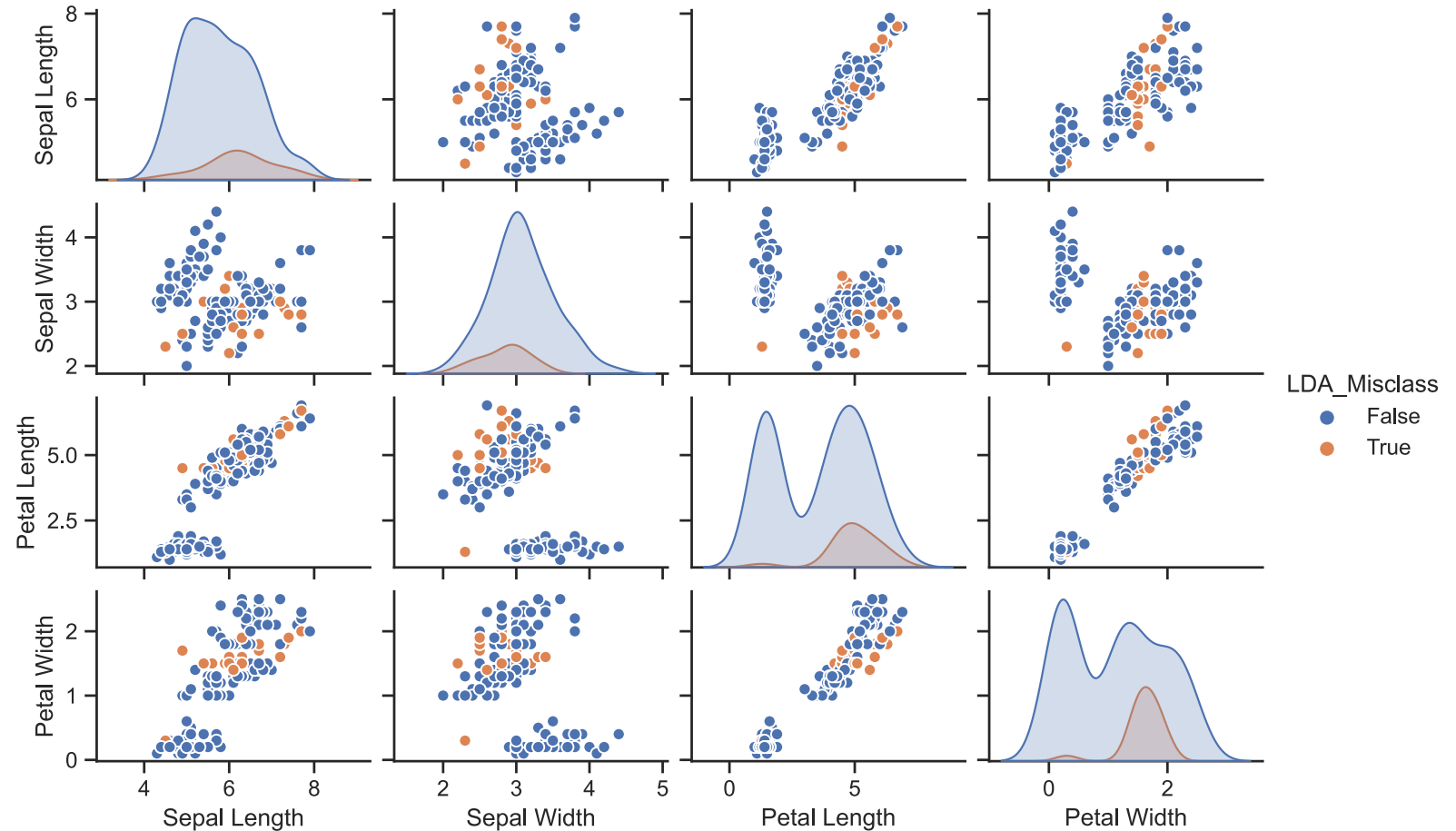
Question: How well do you think this is going to perform on the observed data?

LDA - Irises

Overall Mis-Classification
Rate of .13

Class	LDA MAP Prediction		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	49	1	0
Iris-versicolor	0	42	8
Iris-virginica	0	11	39

LDA is struggling to
correctly classify the
versicolor and virginica
classes...

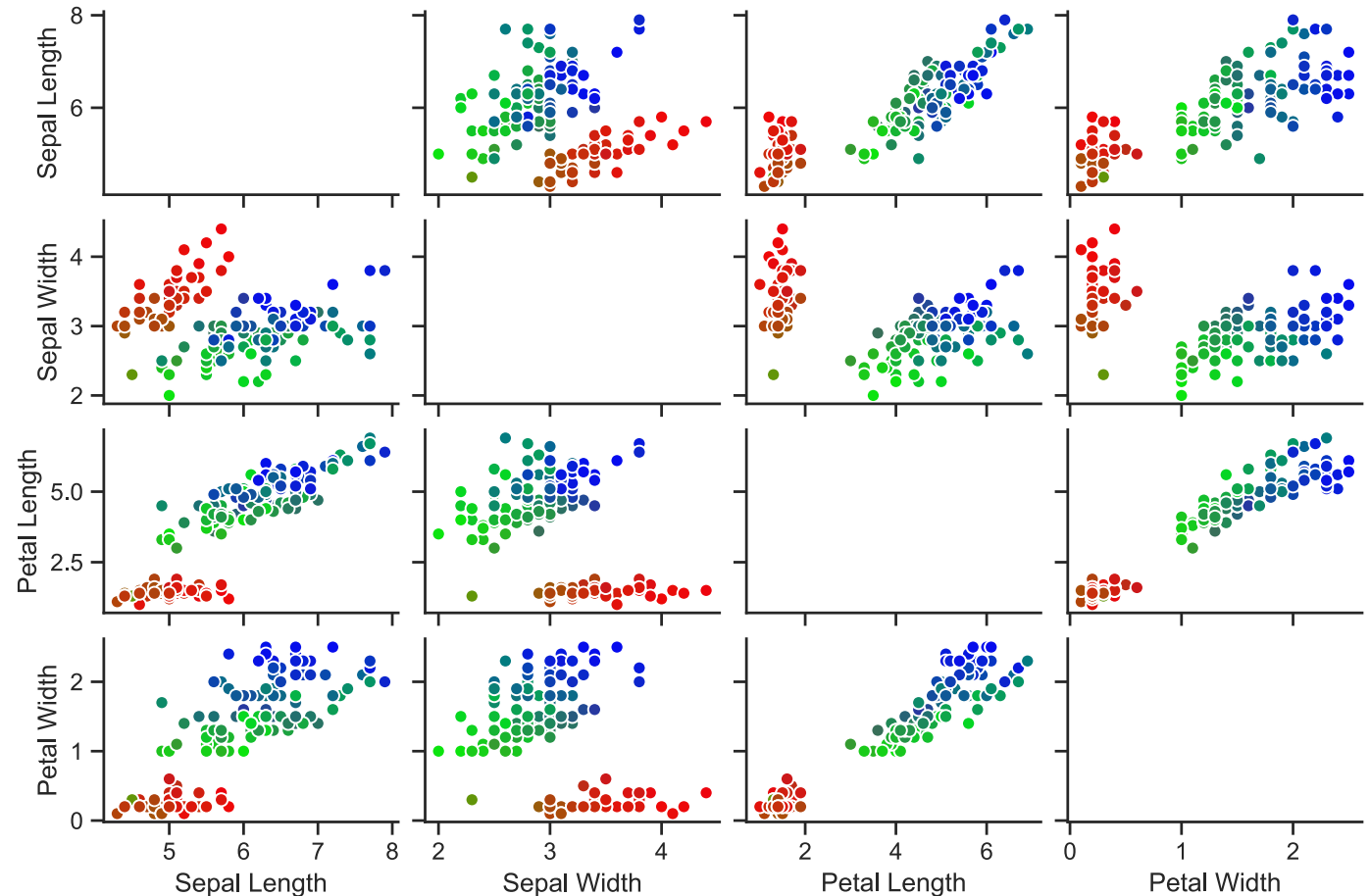


LDA - Irises

Probability Based Coloring

- Red/Green/Blue assigned to the class probabilities.

Note the “fuzzy” boundaries between clusters. Those indicate poor discrimination between types.



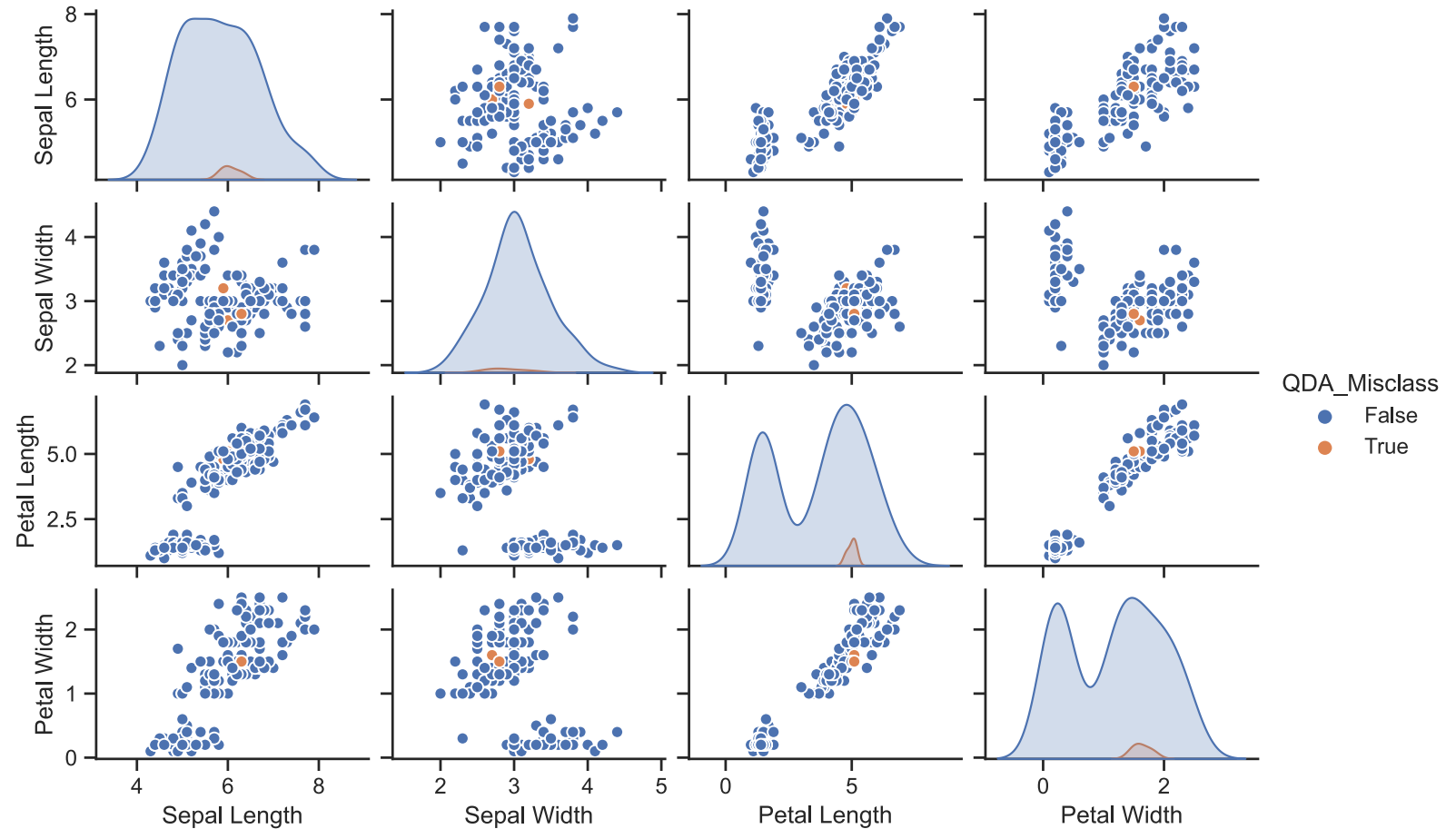
QDA - Irises

QDA just allows Σ to vary between types...

Overall Mis-Classification Rate of .03

Class	QDA MAP Prediction		
	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	50	0	0
Iris-versicolor	0	48	2
Iris-virginica	0	1	49

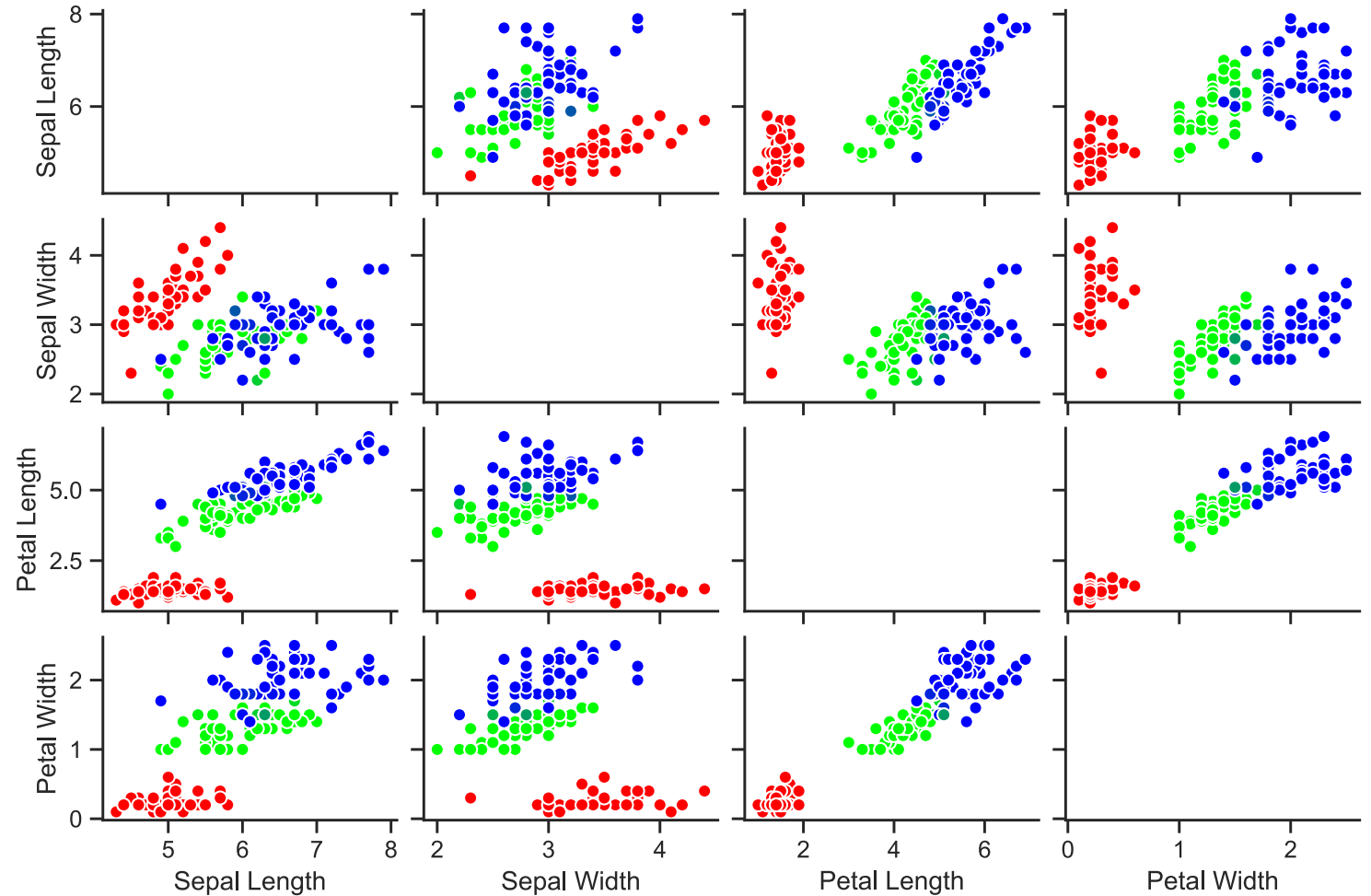
QDA is much better than LDA here.



QDA - Irises

Probability Based
Coloring.

Much sharper
distinctions between iris
types. This classifier
allows for more class
separation because Σ is
allowed to vary



Classification Overview

The general idea of classification is to obtain the probability of class membership and classify each observation with the most probable class.

The devil is in the details though...

- How class is related to the distribution of the continuous predictors is the most important thing.
- We saw that LDA, which assumes each group has the same covariance matrix performs much worse than QDA, which assumes a unique cov matrix for each group.
- However, we might have issues with overfitting! Why should we believe that our sample estimates of group covariance matrices are correct?
- We will address this issue later with hierarchical models.

Note: We did not address changing the prior distributions or the relative costs.

Bayesian Regression

Classification is all well and good when it comes to discrete outcomes, but continuous outcomes need regression!

$$y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

This is the same as:

$$P(y_i | \boldsymbol{\beta}, \sigma_\varepsilon^2) \sim N(\mathbf{X}_i^T \boldsymbol{\beta}, \sigma_\varepsilon^2)$$

Likelihood

There are two groups of parameters $\boldsymbol{\beta}$ and σ_ε^2 that we can use Bayesian approaches on. Today, we will only focus on $\boldsymbol{\beta}$.

NOTE: Even if we let σ_ε^2 be unknown, the expected values of $\boldsymbol{\beta}$ are the same

Bayesian Regression

Before we get into estimation, consider the question of conjugate priors for β ?

$$P(y_i | X_i, \beta, \sigma_\varepsilon^2) \sim N(X_i^T \beta, \sigma_\varepsilon^2)$$

multivariate normal distribution

Two facts (and one assumption) help you determine the conjugate priors.

1. y is normally distributed, its mean is a linear combination of X and β
2. Linear combinations of normally distributed variables is also normal.
3. Let's fix σ_ε^2 as known.

This is a very similar problem to a normal distribution with known variance.

We can use a multivariate normal distribution as a conjugate prior for β !

Bayesian Regression

$$P(y_i | \mathbf{X}_i, \boldsymbol{\beta}, \sigma_\varepsilon^2) \sim N(\mathbf{X}_i^T \boldsymbol{\beta}, \sigma_\varepsilon^2) \quad \text{Likelihood!}$$

$$P(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{Prior!}$$

As the multivariate normal is a conjugate prior for $\boldsymbol{\beta}$, the posterior distribution is also a multivariate normal. Via the power of math....

$$P(\boldsymbol{\beta} | \mathbf{X}_i, y_i, \sigma_\varepsilon^2) \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \quad \text{Posterior!} \quad \text{posteriors are combinations of our priors + our data}$$

$$\boldsymbol{\mu}_\beta = (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1} (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{X}^T \mathbf{y})$$
$$\boldsymbol{\Sigma}_\beta = \sigma_\varepsilon^2 (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1} \quad \text{our covariance matrix}$$

How do these compare with maximum likelihood estimates (frequentist)?

$$\boldsymbol{\mu}_{\beta[OLS]} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$
$$\boldsymbol{\Sigma}_{\beta[OLS]} = \sigma_\varepsilon^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

Bayesian Regression - Ridge

$$\begin{aligned}\boldsymbol{\mu}_\beta &= (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1} (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \mathbf{X}^T \mathbf{y}) \\ \boldsymbol{\Sigma}_\beta &= \sigma_\varepsilon^2 (\boldsymbol{\Sigma}_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1}\end{aligned}$$

As per usual, the EAP estimates ($\boldsymbol{\mu}_\beta$) are a compromise between prior information and the data at hand. By carefully structuring the prior, we can replicate various frequentist regularization approaches.

$$\boldsymbol{\beta} \sim N(0, \sigma_0^2 \mathbf{I})$$

$$\boldsymbol{\mu}_{\beta[Ridge]} = \left(\frac{\sigma_\varepsilon^2}{\sigma_0^2} \mathbf{I} + \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

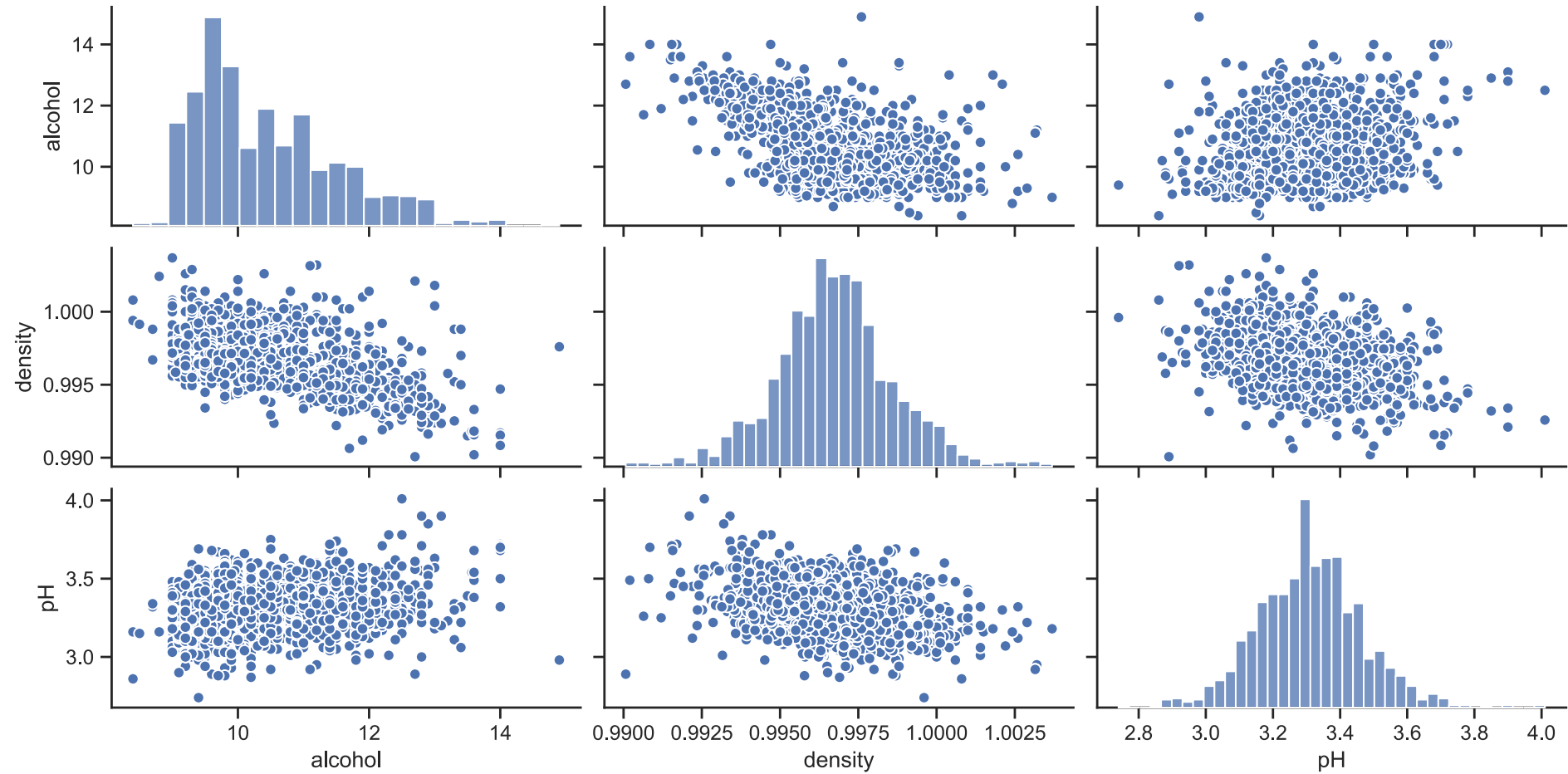
Red Wine Statistics

There is a famous [dataset](#) regarding wine quality as a function of various characteristics. (You will be looking at how to classify white wines in HW2).

Task: Predict alcohol content as a function of density and pH!



Red Wine Statistics



Red Wine Statistics

MLE Estimates for the intercept and coefficients for density and pH:

- Intercept: 280.88
- Density: -272.28
- pH: .28

Strategy: Using a Bayesian Ridge Regression $p(\boldsymbol{\beta}) \sim N(0, \lambda \mathbf{I})$ explore how varying λ impacts the EAP estimates of the regression coefficients



EAP Estimates

