Machine Learning Department
School of Computer Science
Carnegie Mellon University

Gaussian Naïve Bayes

Naïve Bayes Readings:

"Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression" (Mitchell, 2016)

Murphy 3
Bishop --
HTF --
Mitchell 6.1-6.10

Optimization Readings: (next lecture)

Lecture notes from 10-600 (see Piazza note)

"Convex Optimization" Boyd and Vandenberghe (2009) [See Chapter 9. This advanced reading is entirely optional.]

Matt Gormley
Lecture 6
February 6, 2016

Reminders

- **Homework 2: Naive Bayes**
  - Release: Wed, Feb. 1
  - Due: Mon, Feb. 13 at 5:30pm
- **Homework 3: Linear / Logistic Regression**
  - Release: Mon, Feb. 13
  - Due: Wed, Feb. 22 at 5:30pm

Naïve Bayes Outline

- **Probabilistic (Generative) View of Classification**
  - Decision rule for probability model
- **Real-world Dataset**
  - Economist vs. Onion articles
  - Document → bag-of-words → binary feature vector
- **Naive Bayes: Model**
  - Generating synthetic "labeled documents"
  - Definition of model
  - Naive Bayes assumption
  - Counting # of parameters with / without NB assumption
- **Naïve Bayes: Learning from Data**
  - Data likelihood
  - MLE for Naive Bayes
  - MAP for Naive Bayes
- **Visualizing Gaussian Naive Bayes**

Last Lecture

This Lecture

Naive Bayes: Model

Whiteboard

- Generating synthetic "labeled documents"
- Definition of model
- Naive Bayes assumption
- Counting # of parameters with / without NB assumption

What’s wrong with the Naïve Bayes Assumption?

The features might not be independent!!

- Example 1:
  - If a document contains the word “Donald”, it’s extremely likely to contain the word “Trump”
  - These are not independent!

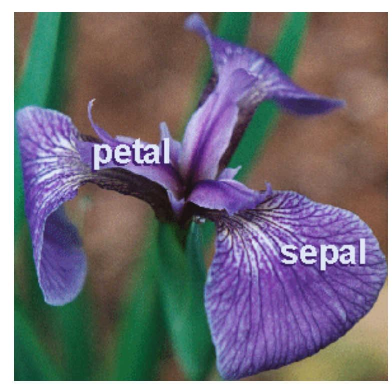
★ ELECTION 2016 ★ MORE ELECTION COVERAGE

Trump Spends Entire Classified National Security Briefing Asking About Egyptian Mummies

NEWS IN BRIEF August 18, 2016

VOL 52 ISSUE 32 · Politics · Politicians · Election 2016 · Donald Trump

- Example 2:
  - If the petal width is very high, the petal length is also likely to be very high



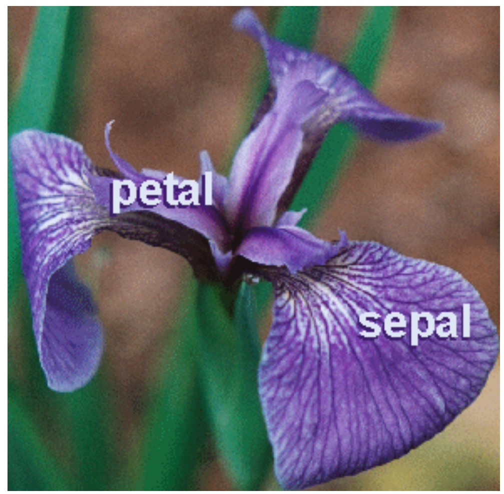
Naïve Bayes: Learning from Data

Whiteboard

- Data likelihood
- MLE for Naive Bayes
- MAP for Naive Bayes

VISUALIZING NAÏVE BAYES

Slides in this section from William Cohen (10-601B, Spring 2016)

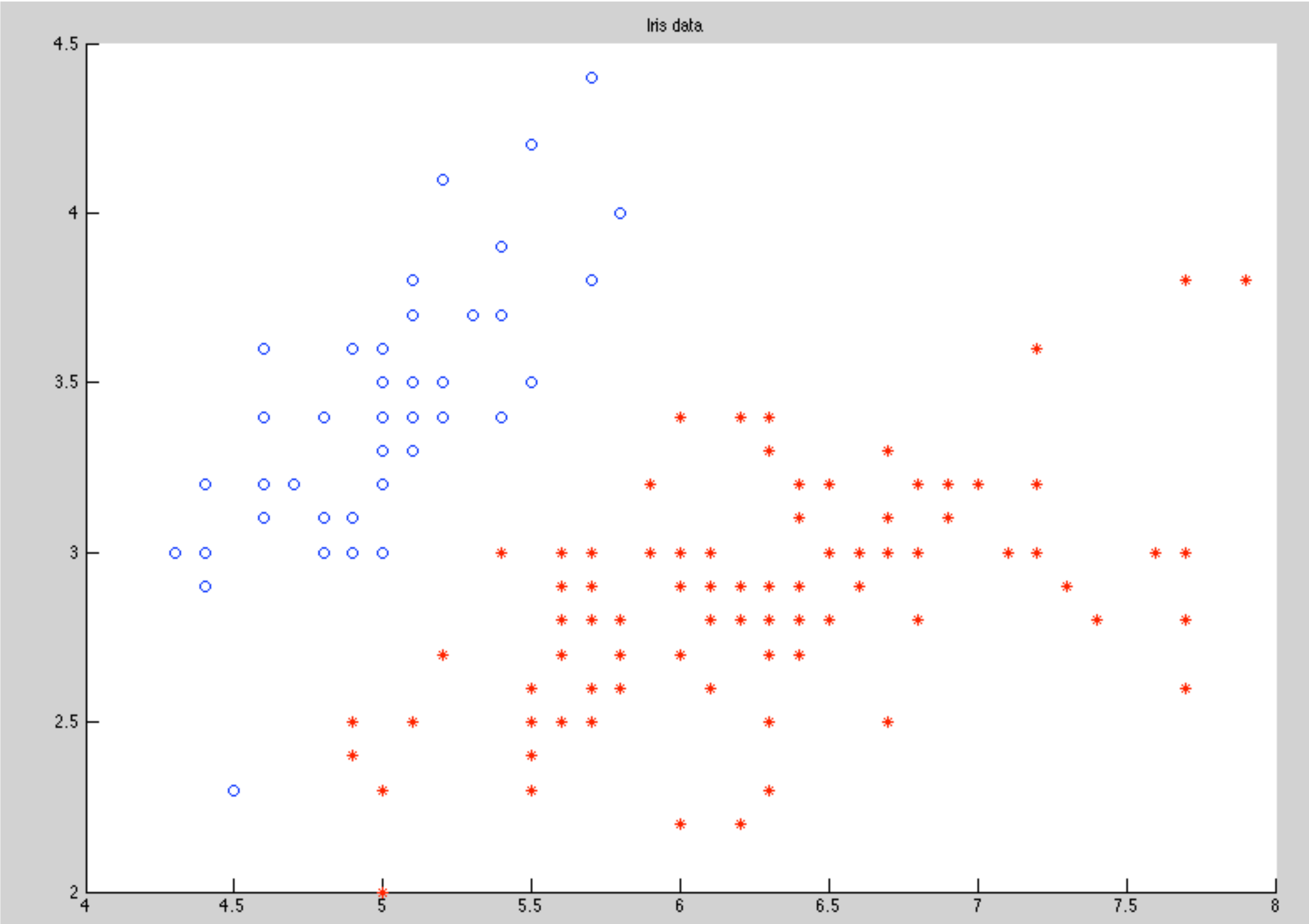

petal
sepal

Fisher Iris Dataset

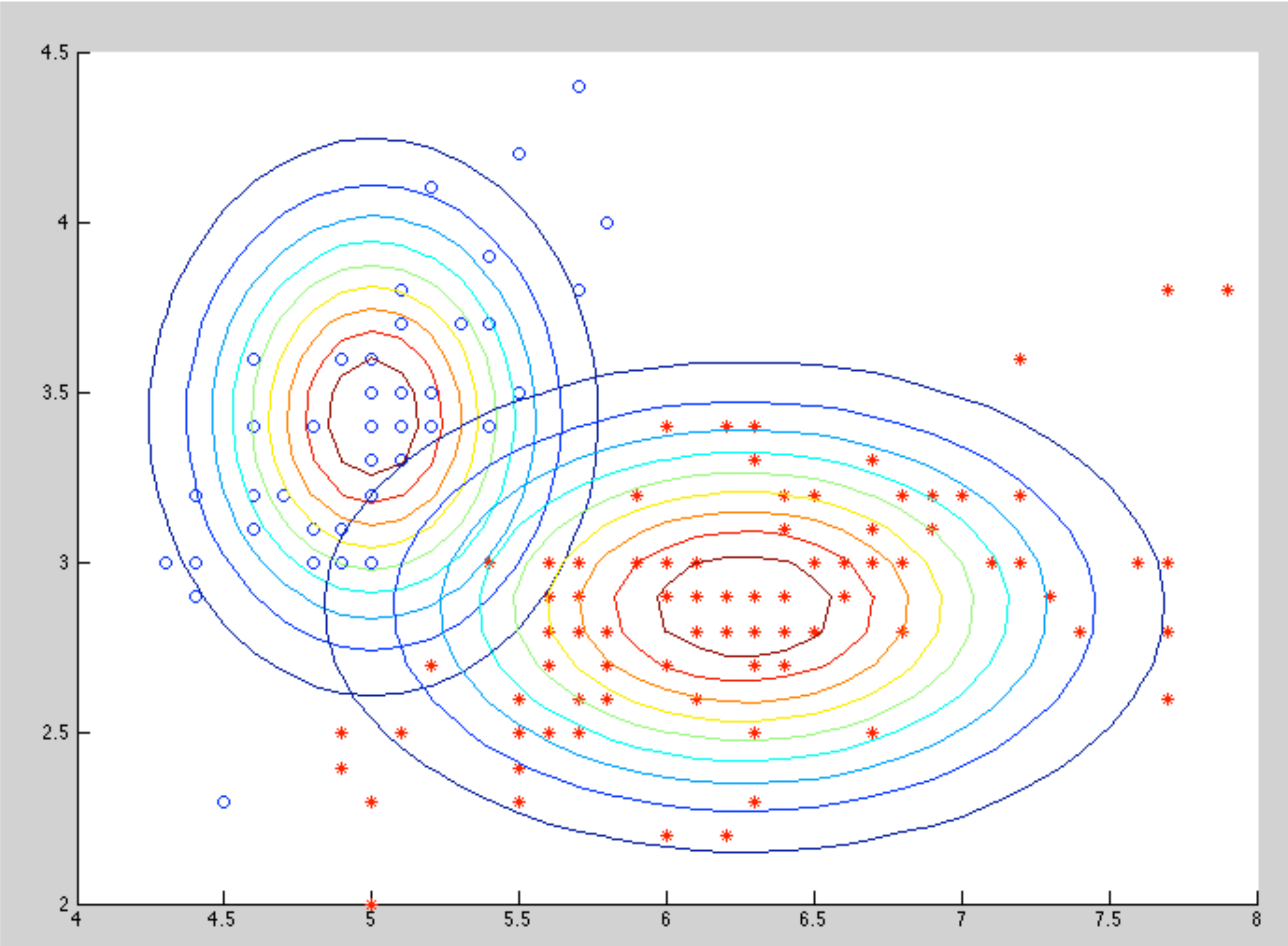
Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	6.7	3.0	5.0	1.7

Full dataset: https://en.wikipedia.org/wiki/Iris_flower_data_set



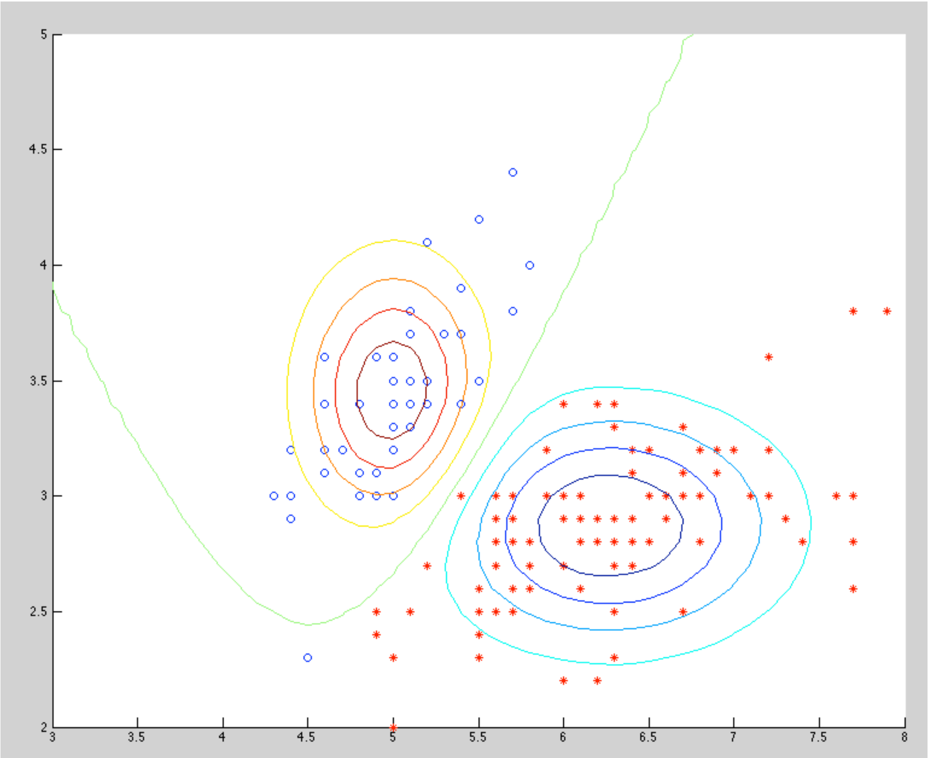
Slide from William Cohen

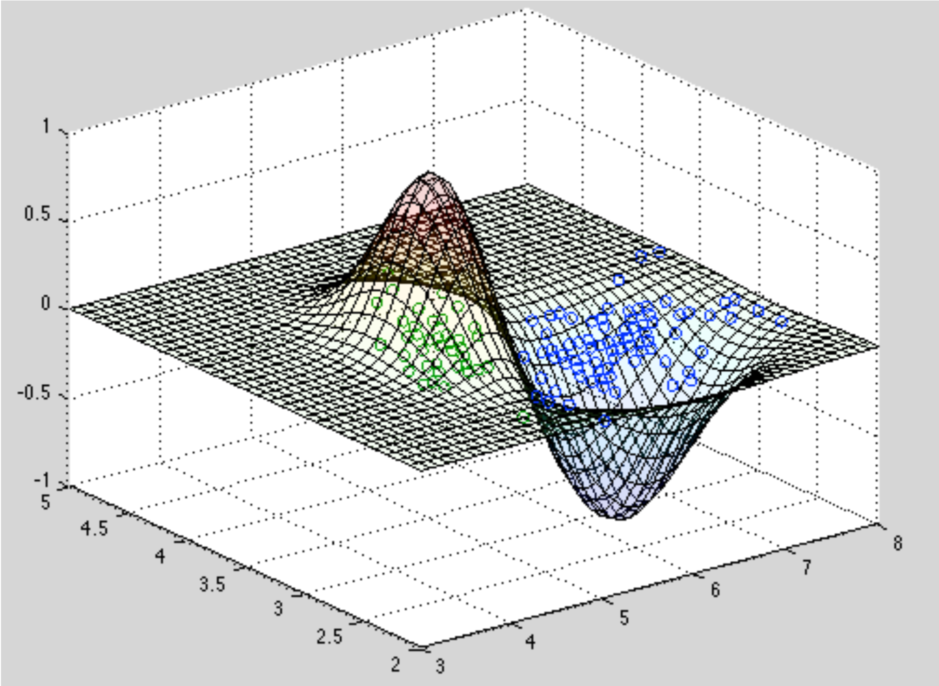


Slide from William Cohen

Plot the difference of the probabilities

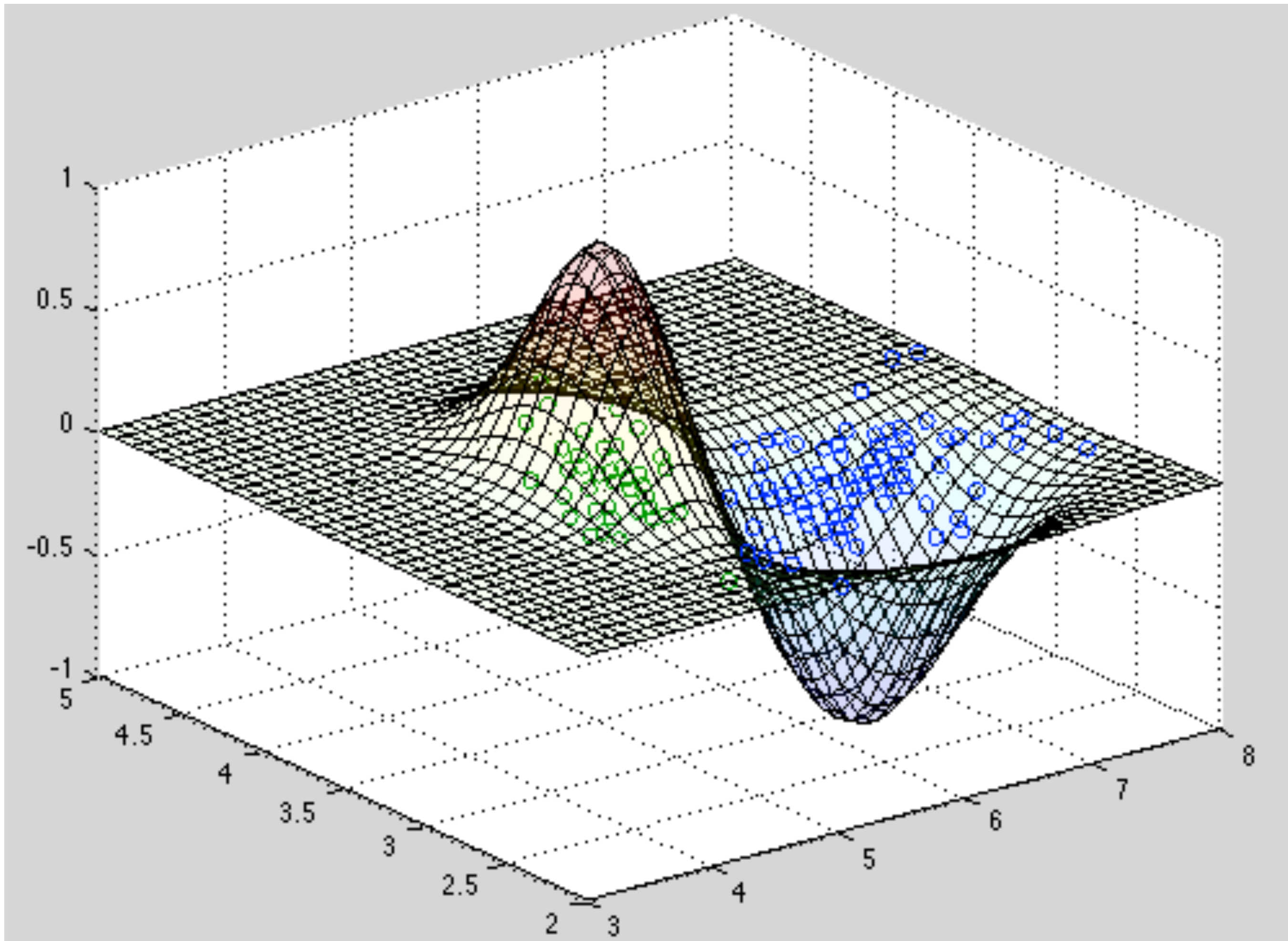
z-axis is the difference of the posterior probabilities: $p(y=1 \mid \mathbf{x}) - p(y=0 \mid \mathbf{x})$



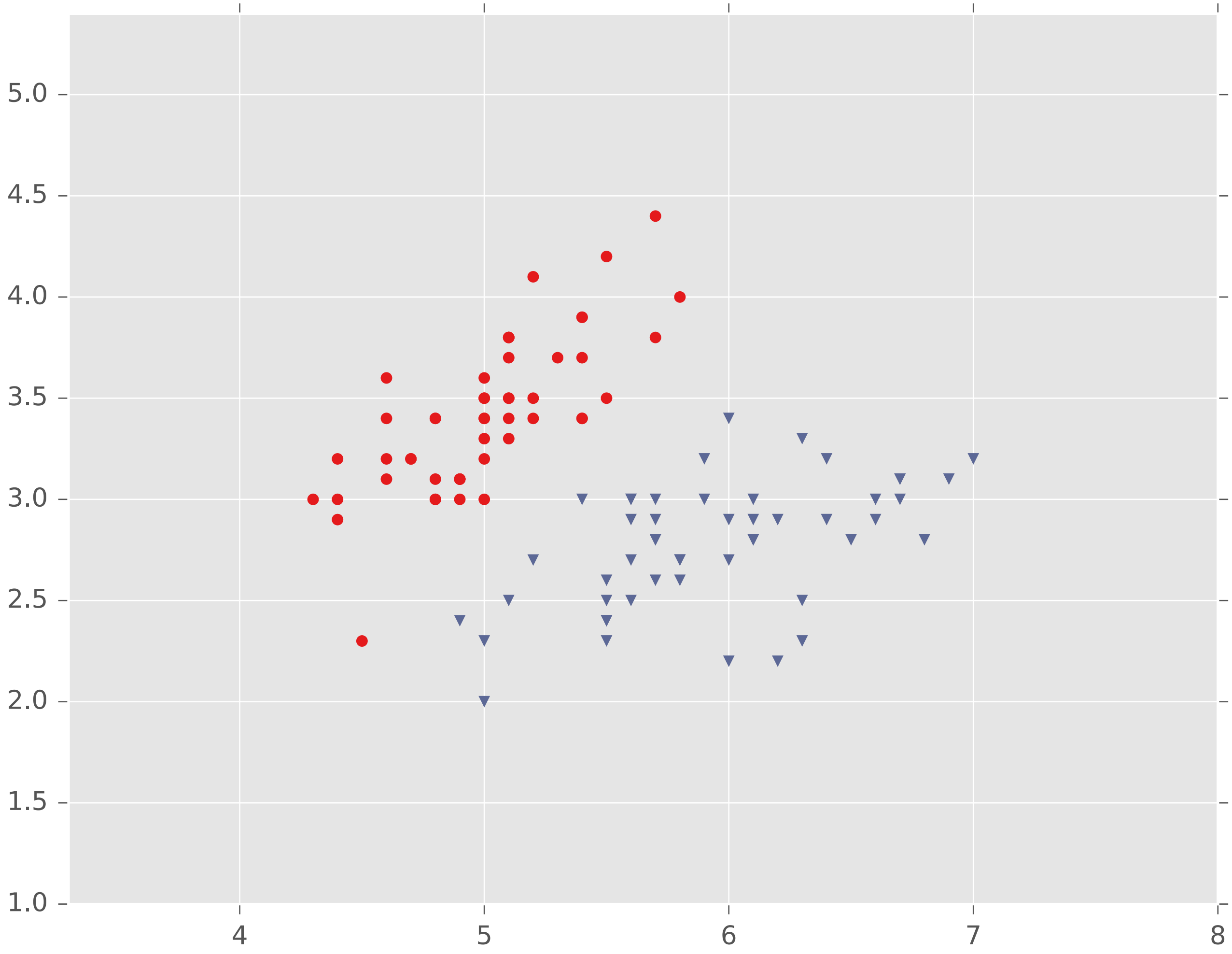


Slide from William Cohen

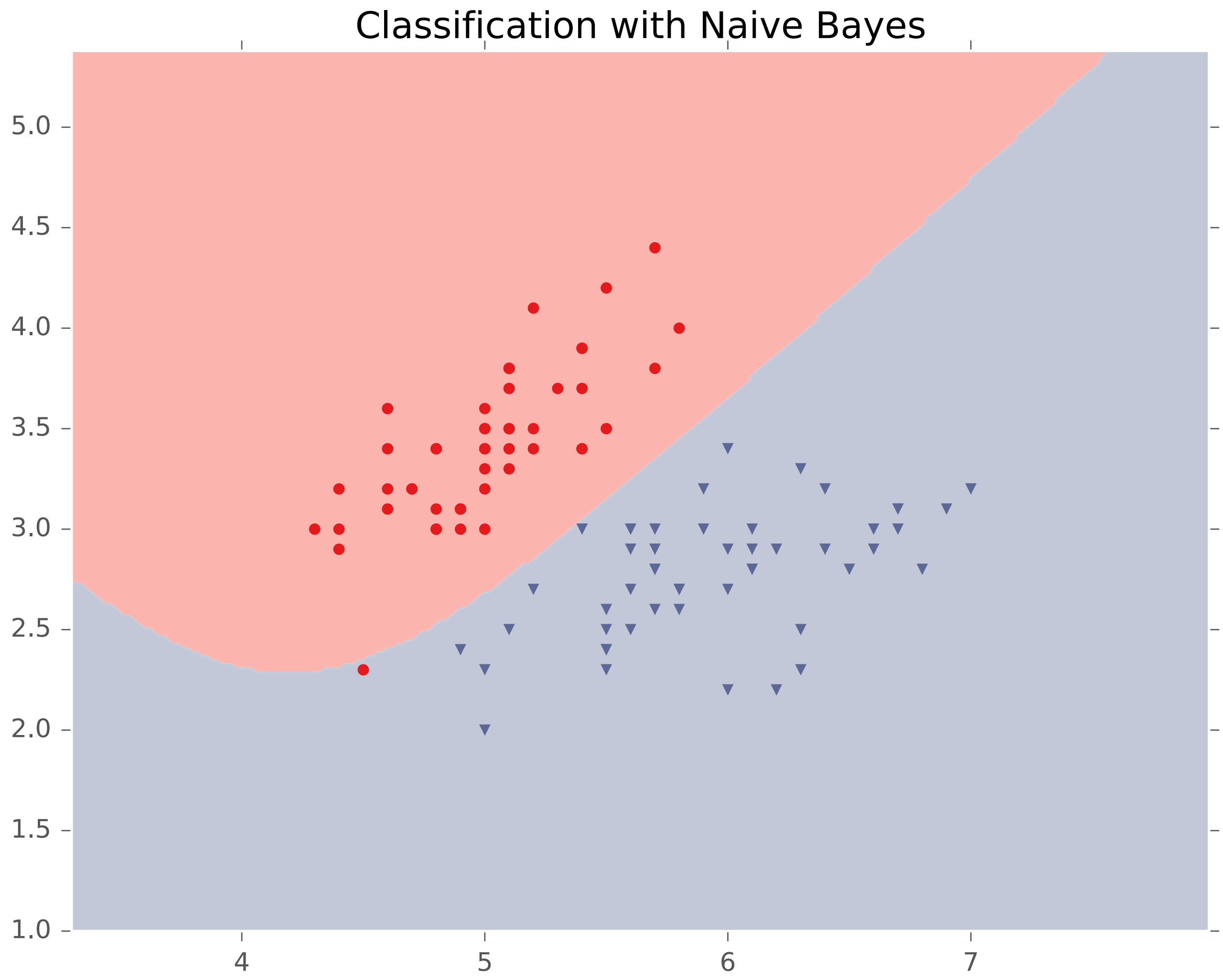
Question: what does the *boundary* between positive and negative look like for Naïve Bayes?



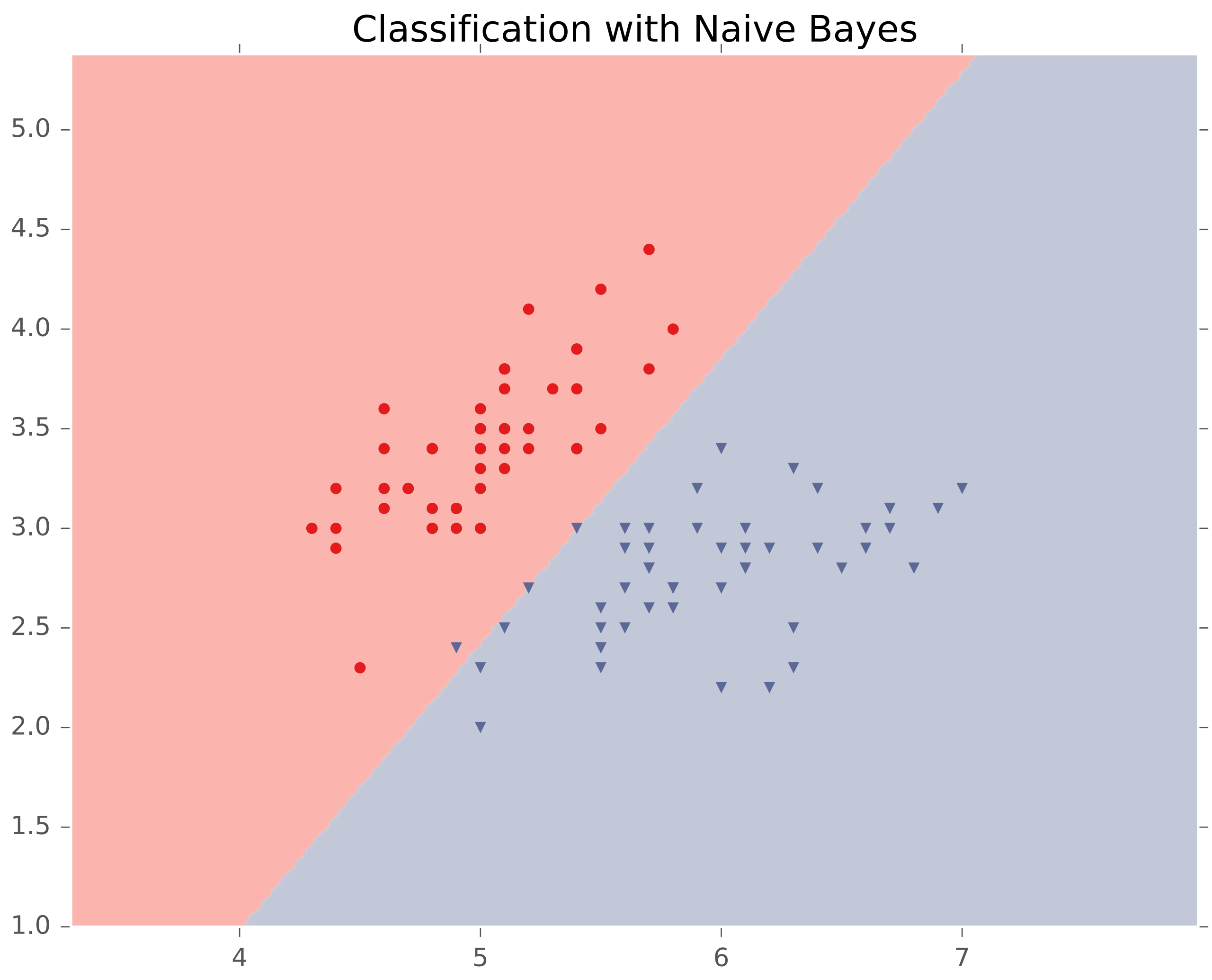
Iris Data (2 classes)



Iris Data (sigma not shared)

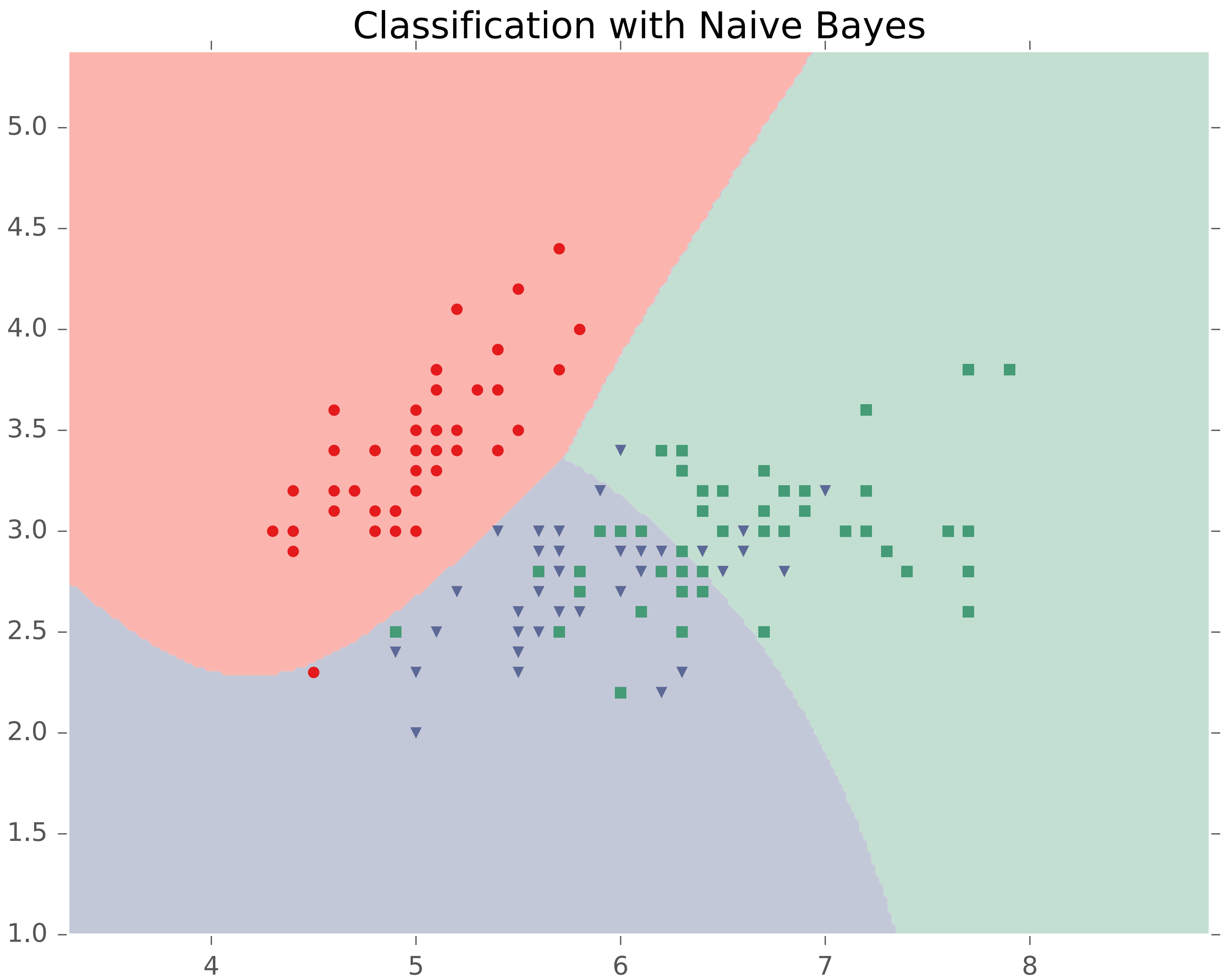


Iris Data (sigma=1)

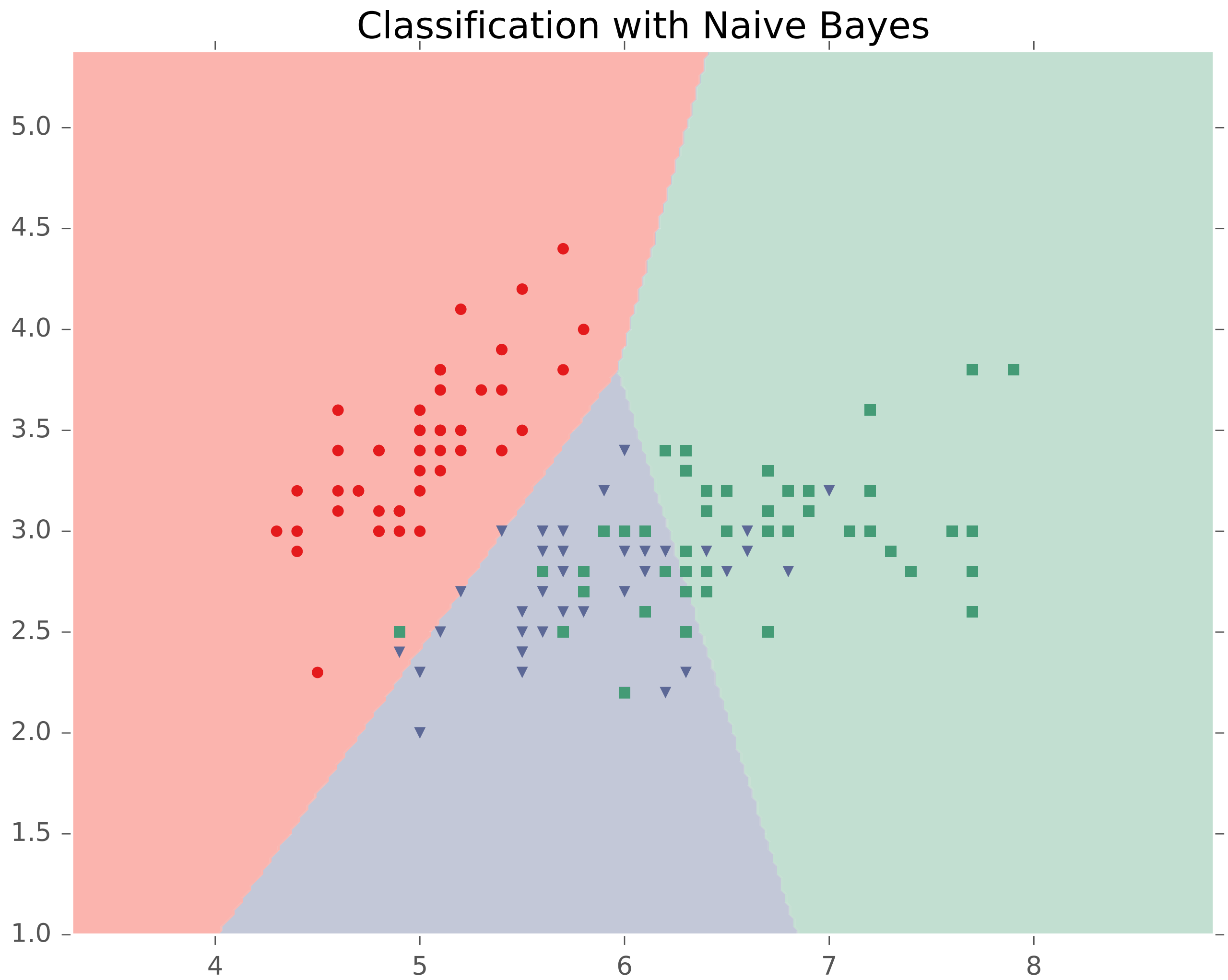


Iris Data (3 classes)

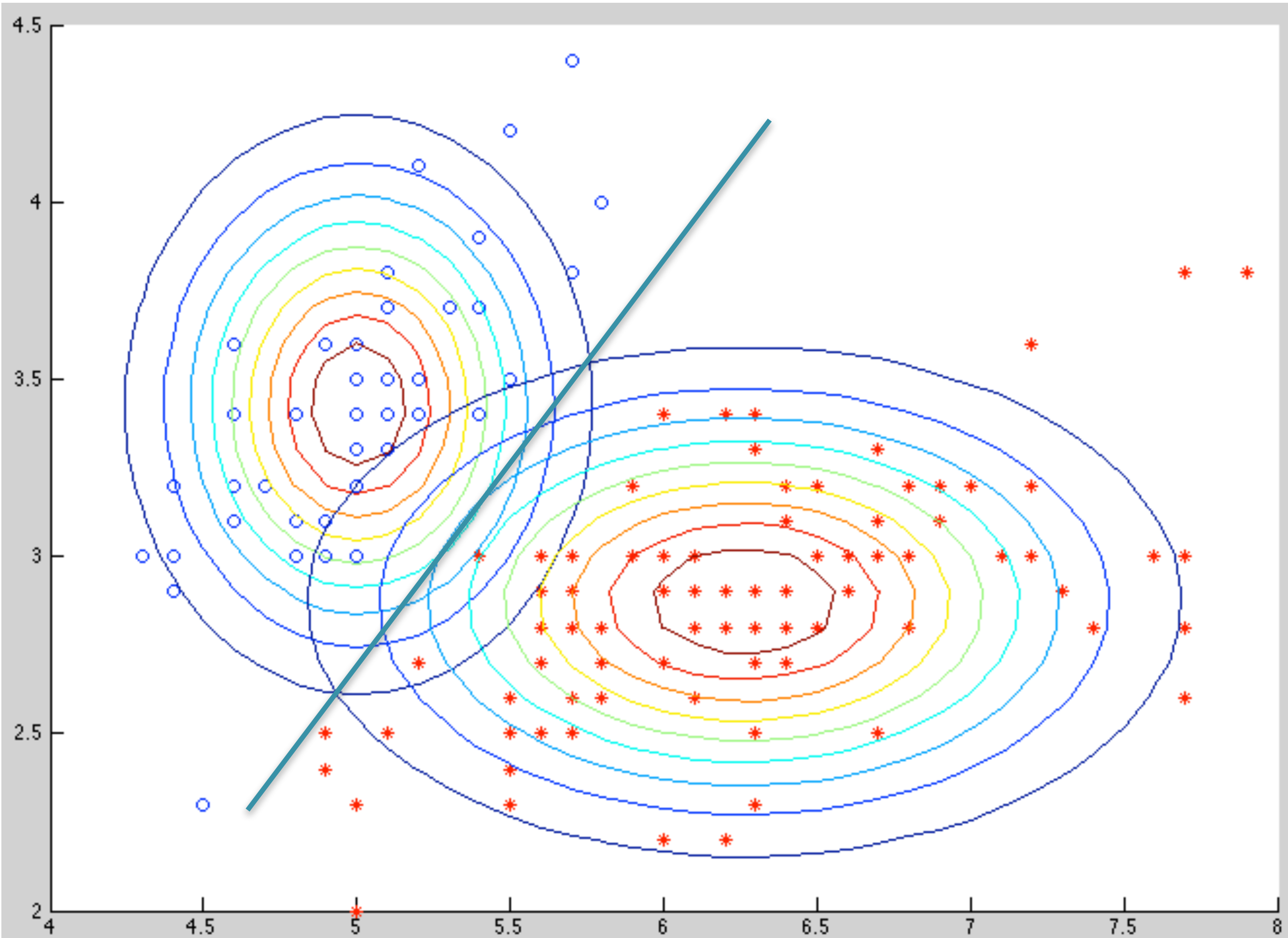
Iris Data (sigma not shared)



Iris Data (sigma=1)



Naïve Bayes has a **linear** decision boundary (if sigma is shared across classes)



Slide from William Cohen (10-601B, Spring 2016)

Figure from William Cohen (10-601B, Spring 2016)

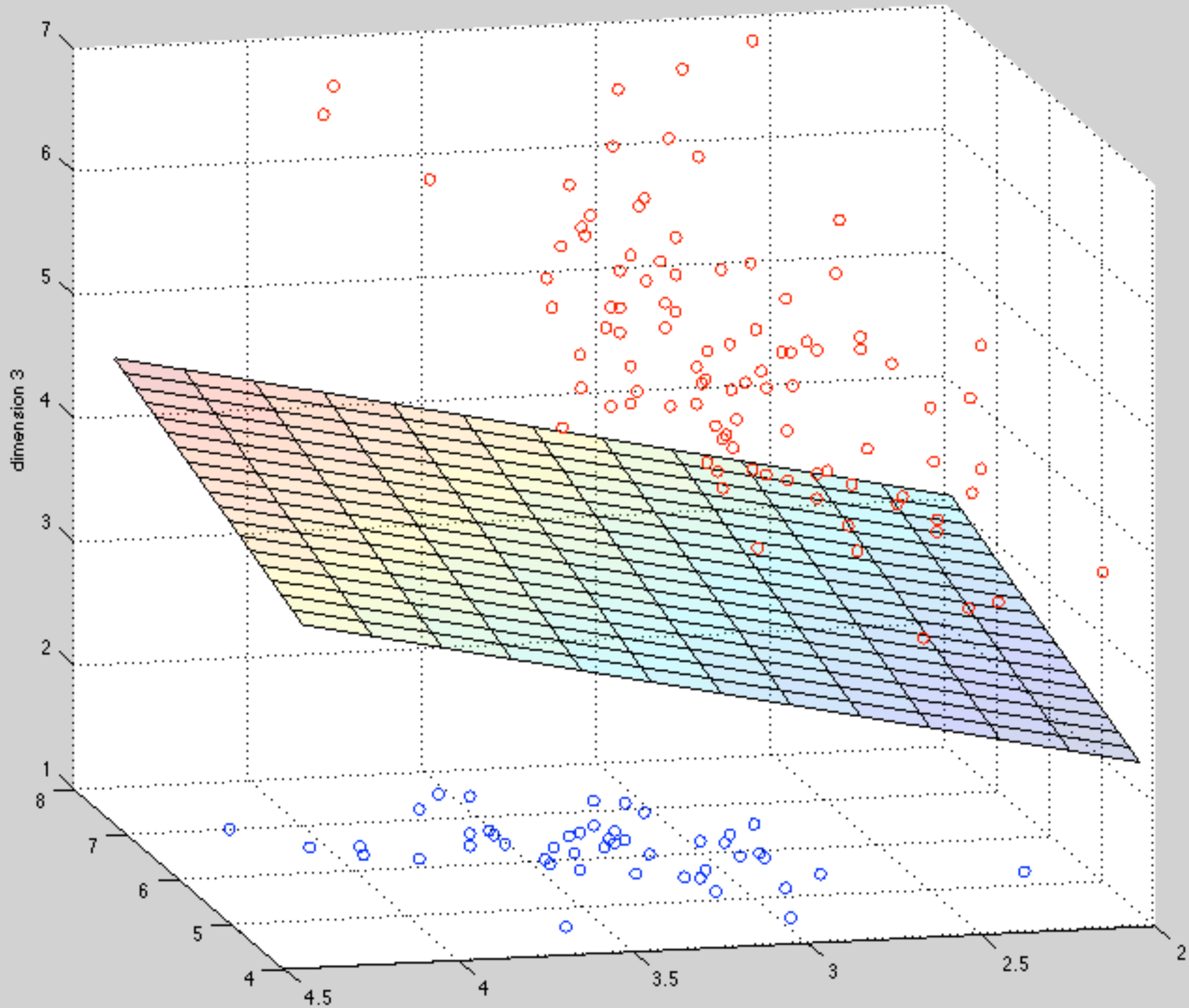
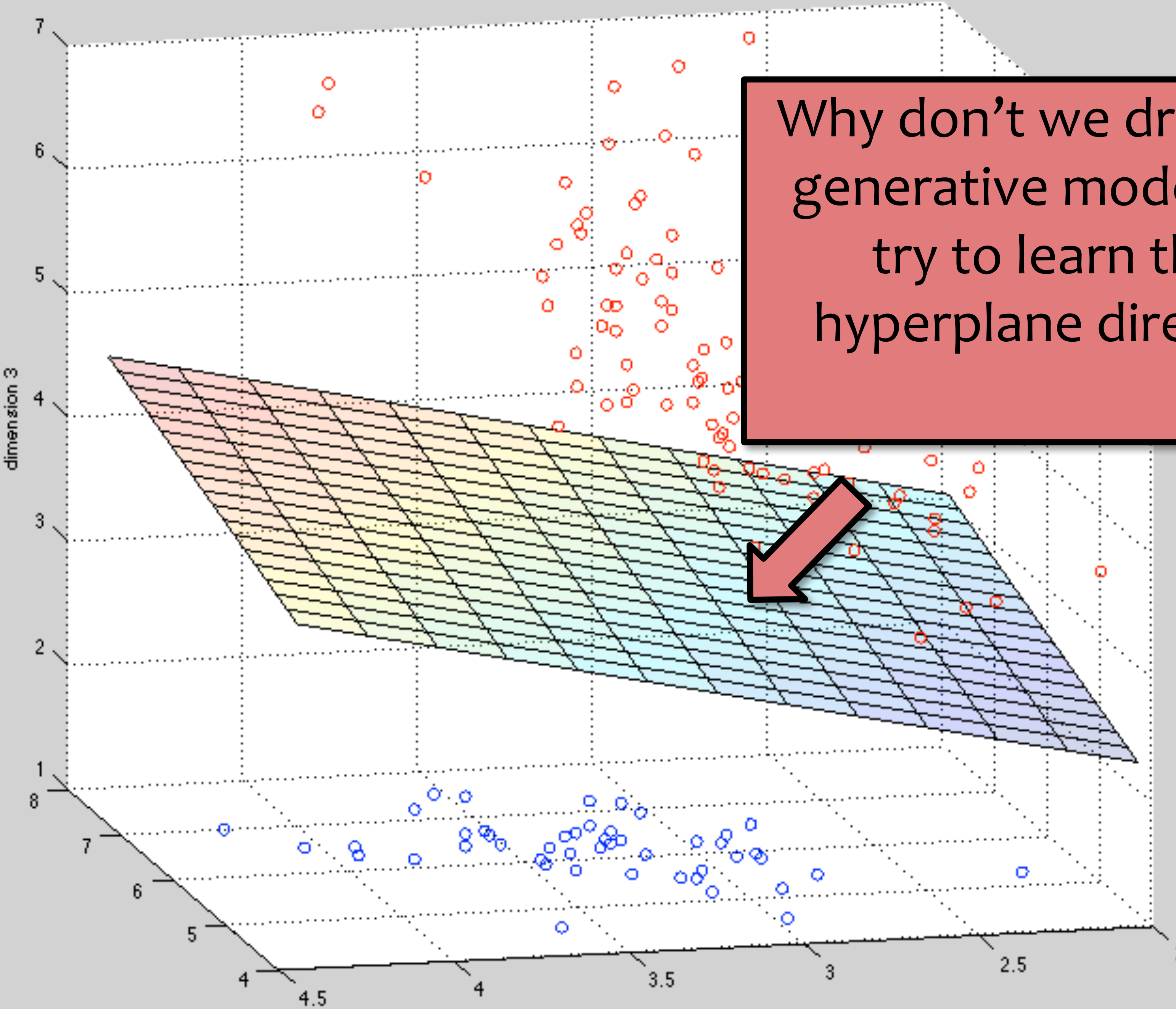


Figure from William Cohen (10-601B, Spring 2016)



Why don't we drop the generative model and try to learn this hyperplane directly?

Beyond the Scope of this Lecture

- **Multinomial** Naïve Bayes can be used for **integer features**
- **Multi-class** Naïve Bayes can be used if your classification problem has > 2 classes

Summary

1. Naïve Bayes provides a framework for **generative modeling**
2. Choose $p(x_m \mid y)$ appropriate to the data (e.g. Bernoulli for binary features, Gaussian for continuous features)
3. Train by **MLE** or **MAP**
4. Classify by maximizing the posterior