Supervised Machine Learning and Learning Theory

Lecture 4: Classification

September 17, 2024

In-class quiz questions

- Can you explain the three terms in the bias-variance trade-off?
- Is the bias-variance trade-off specific to linear regression, or does it apply to any machine learning model?
- For K-nearest neighbors, does the model become more flexible or less flexible for a larger K?
- Suppose you would like to fit a polynomial function, how would you do that using linear regression?

In-class quiz questions

- What is the relation between the gradient, the derivatives, and the partial derivatives?

- For a matrix X, is the rank of X the same as its transpose $X^\top$? Why or why not?

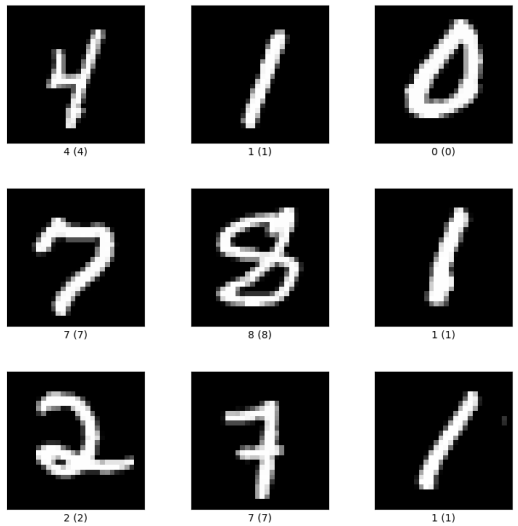
- Name several commonly used metrics for linear regression?

Lecture plan

- **Logistic regression**

Classification example I

- Handwritten digit classification
- Colored handwritten digits
- Street view house numbers



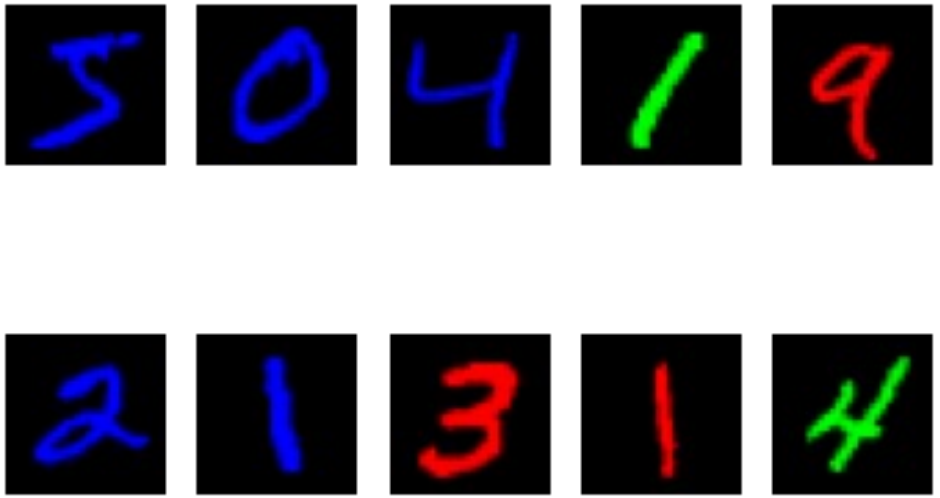
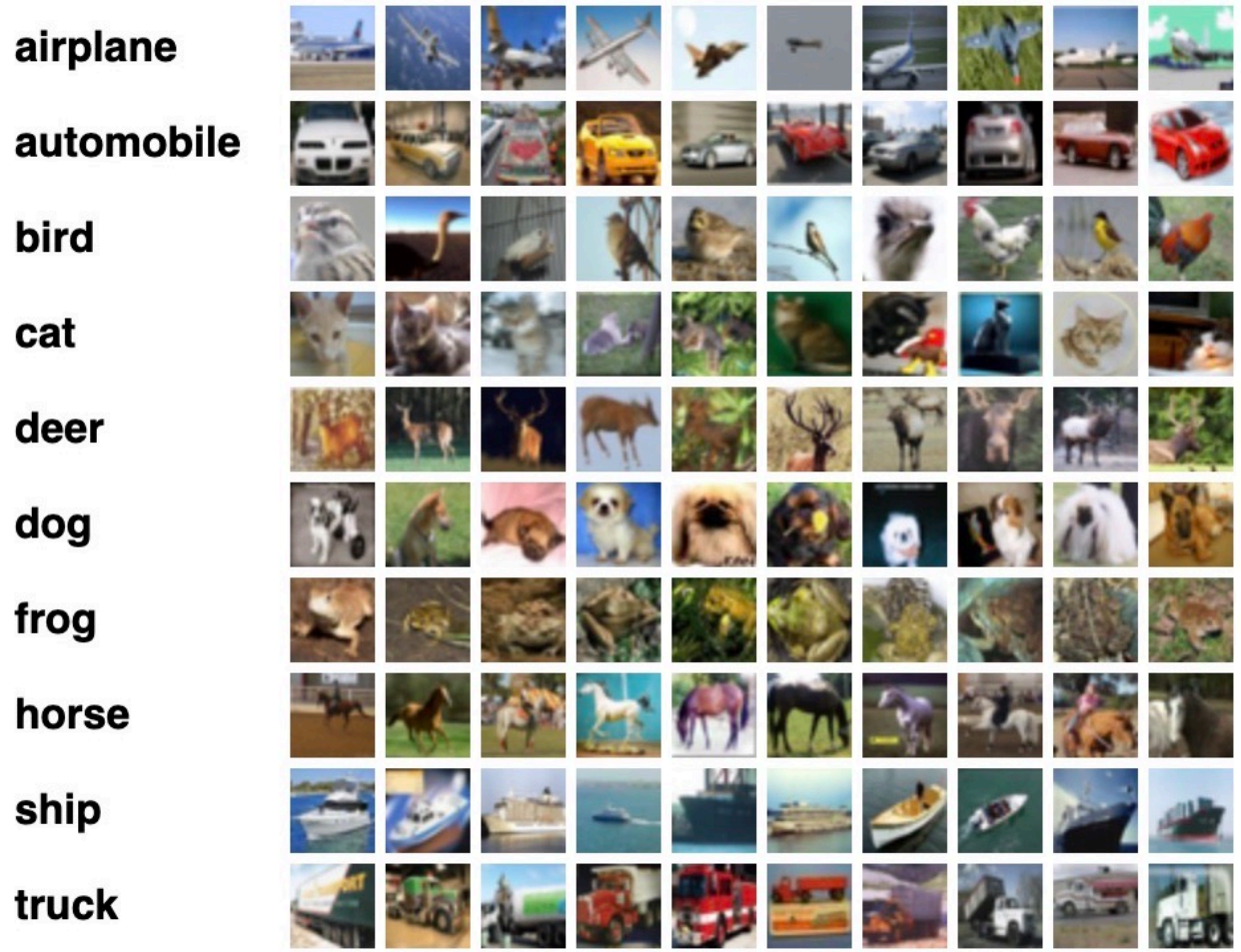


Image classification

- Image classification: assign a label to an entire image or photograph
- Object recognition



CIFAR-10: Canadian Institute For Advanced Research 60,000 images in 10 different classes, with 6,000 images of each class

A classification problem

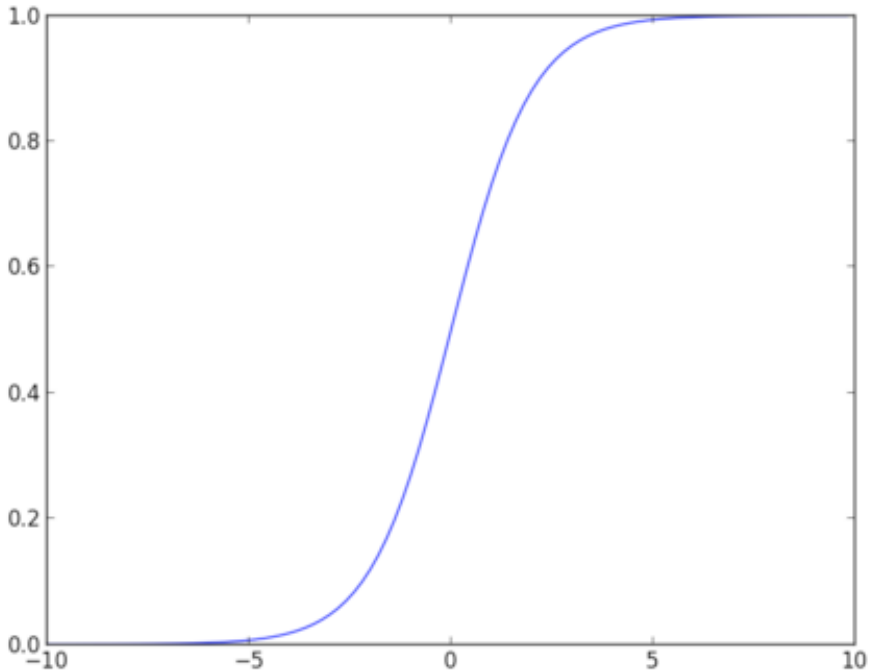
- A dataset containing information on ten thousand customers
  - **default**: whether the customer defaulted on their debt
  - **student**: whether the customer is a student
  - **balance**: the average balance that the customer has remaining on their credit card after making their monthly payment
  - **income**: income of customer

- Predict which customers will default on their credit card debt

```
## # A tibble: 10,000 x 4
##    default student balance income
##    <fct>   <fct>     <dbl>  <dbl>
##  1 No      No         730. 44362.
##  2 No      Yes        817. 12106.
##  3 No      No        1074. 31767.
##  4 No      No         529. 35704.
##  5 No      No         786. 38463.
##  6 No      Yes        920.  7492.
##  7 No      No         826. 24905.
##  8 No      Yes        809. 17600.
##  9 No      No        1161. 37469.
## 10 No      No           0  29275.
## # … with 9,990 more rows
```

The logistic loss

- While zero-one loss makes intuitive sense, minimizing the zero-one loss is computationally hard
- Logistic loss provides an approximation of the zero-one loss
- The logistic function: $\frac{1}{1+e^{-v}}$ (note: sigmoid function ranges between $-1,1$)



Understanding the log loss

- Odds function: (we'll insert the value of v from a linear model)

$$\frac{\exp(v)}{1+\exp(v)}, \text{ ranges between zero and one}$$

- Logistic loss: negative log of the logistic function

$$\ell(v) = -\log\frac{\exp(v)}{1+\exp(v)} = \log(1+\exp(-v))$$

- Exercise: What is the value of $\ell(10)$, and what about $\ell(-10)$?

Using logistic regression for binary classification

- Suppose the labels are either $+1$ or -1. For every sample x_i, y_i, suppose x_i includes p features in total, indexed by subscripts as $x_{i,1}, x_{i,2}, \ldots, x_{i,p}$
- Coefficients of the logistic regression model: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Let

$$v_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$

- The log-loss of x_i, y_i is

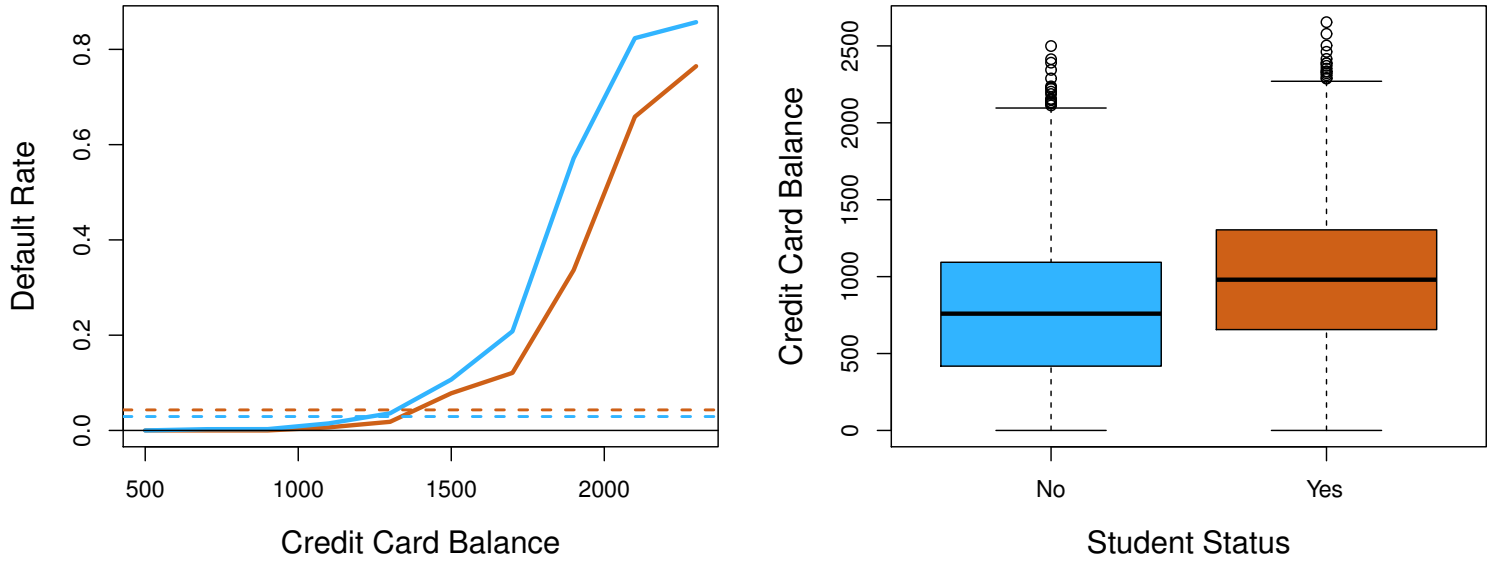
$$\log(1 + \exp(-y_i \cdot v_i))$$

- The averaged log-loss applied to a training set of size n is

$$\frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \cdot v_i))$$

Example: Predicting credit card default

- Customers with a high balance are more likely to default
- Students tend to have higher balances
- Among customers with a given balance, students are less likely to default



- **Question:** How can we use $(\text{student}, \text{balance}, \text{income})$ to predict default?

Logit model

- The logit model of a sample X, Y (log-odds) is

$$\log\left[\frac{Pr(Y=1|X)}{Pr(Y=0|X)}\right] = \beta^\top X = \beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income},$$

where $X = (1, student, balance, income)$, and this is the same as:

$$\Pr[Y=1|X] = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{student} + \beta_2 \cdot \text{balance} + \beta_3 \cdot \text{income})}}$$

- Find $\beta_0, \beta_1, \beta_2, \beta_3$ by minimizing the averaged log-loss over the training set. Exercise: write down the training loss for this example

- Prediction: if $v_i > 0, \hat{y}_i = +1$; if $v_i \leq 0, \hat{y}_i = -1$

Maximum likelihood estimation

- **Maximum likelihood estimation (MLE):** likelihood of Y given X

$$\Pr[Y = 1|X] = \frac{e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}$$

$$\Pr[Y = 0|X] = \frac{1}{1 + e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}$$

- **Objective of MLE:** Find the values of β that maximizes the likelihood of observing n sample points

Maximum likelihood estimation

- The likelihood of the training data is the overall probability for a fixed set of coefficients $\beta_0, \cdots, \beta_p$ (the case of binary labels 0,1). Here we are taking the **product** of all the individual probabilities from n samples:

$$\prod_{i=1}^{n} Pr(Y = y_i | X = x_i) = \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots\beta_p x_{ip}}} \prod_{j:y_j=0} \frac{1}{1+e^{\beta_0+\beta_1 x_{j1}+\cdots\beta_p x_{jp}}}$$

Probability of response = 1 | Probability of response = 0

Assume samples are independent

- Easier to compute after taking negative log of the above likelihood. Why?

MLE for multi-class classification

- Suppose the label Y takes values in $\{1,2,\cdots,K\}$. Let X be a p-dimensional feature vector
- Let u be a vector such that we set a linear model for each class

$$u_1 = \beta_{0,1} + \beta_{1,1}X_1 + \cdots \beta_{p,1}X_p$$
$$\cdots$$
$$u_K = \beta_{0,K} + \beta_{1,K}X_1 + \cdots \beta_{p,K}X_p$$

- Cross-entropy loss

$$\ell(X,Y) = -\log \frac{\exp(u_Y)}{\sum_{i=1}^{K} \exp(u_i)}$$

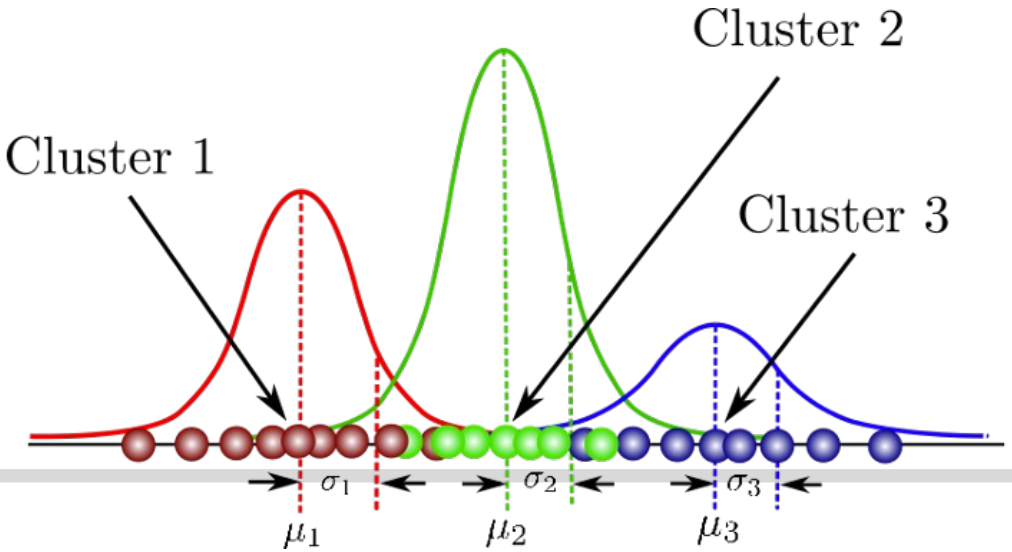
- Verify that the cross-entropy loss is always positive?

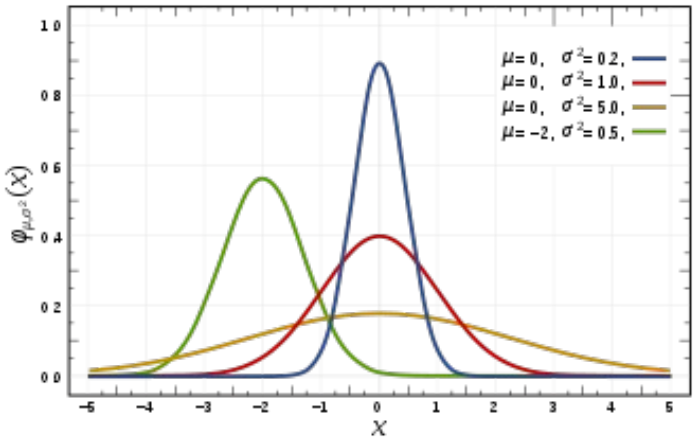
Lecture plan

- **Linear discriminant analysis**

Example

- **Mixture of Gaussians in crabs**: A zoologist considered a dataset of crab measurements among 1000 crabs (during their 1892 Easter vacation). All but one attribute follows a single normal distribution: **Forehead to body length ratio**
- $N(\mu, \sigma^2)$ with mean μ and variance σ^2
- How do we fit a dataset with a mixture of Gaussians?





Reference: http://blog.mrtz.org/2014/04/22/pearsons-polynomial.html

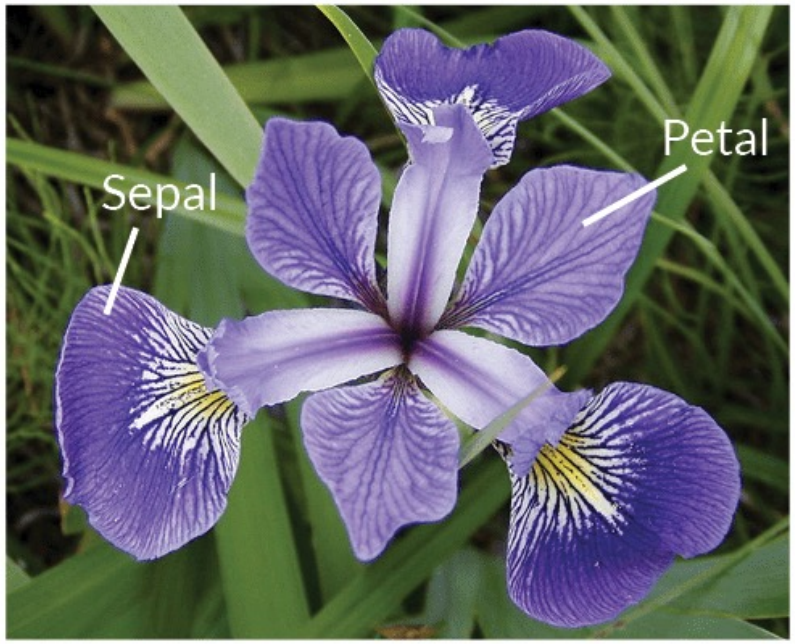
Linear discriminant analysis

Linear Discriminant Analysis (LDA): Suppose we have K classes, we approximate the data distribution of each class with a Gaussian distribution

- π_k: prior probability that a randomly chosen observation comes from the k-th class
- $f_k(X) = \Pr(X|Y = k)$: density function of X coming from the k-th class
- $\Pr(Y = k|X = x)$: probability of x having label k

Example: Iris dataset

- Pattern recognition: Predict class of iris plant. There are three classes

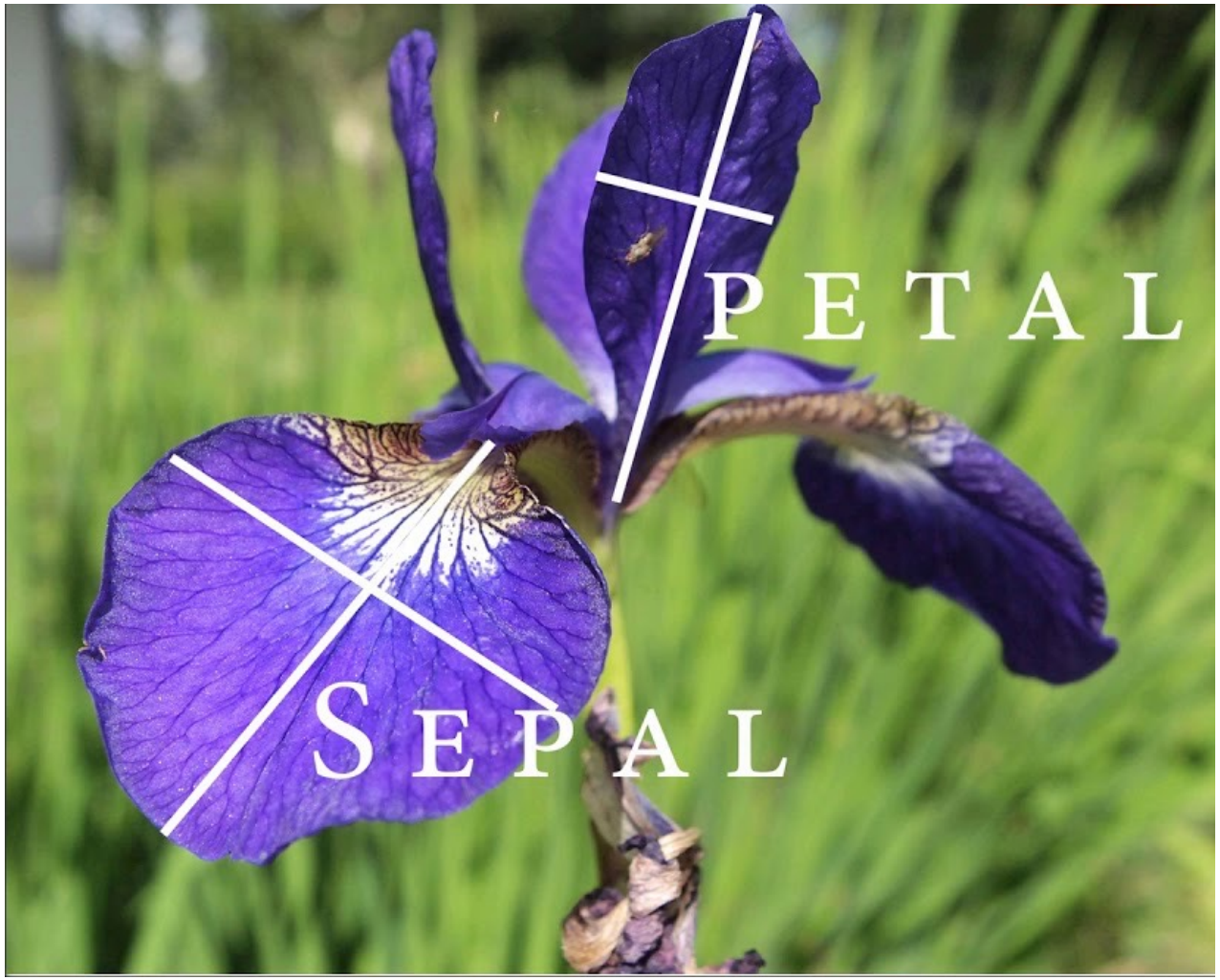


Iris Versicolor

Iris Setosa

Iris Virginica

Sepal and petal of iris

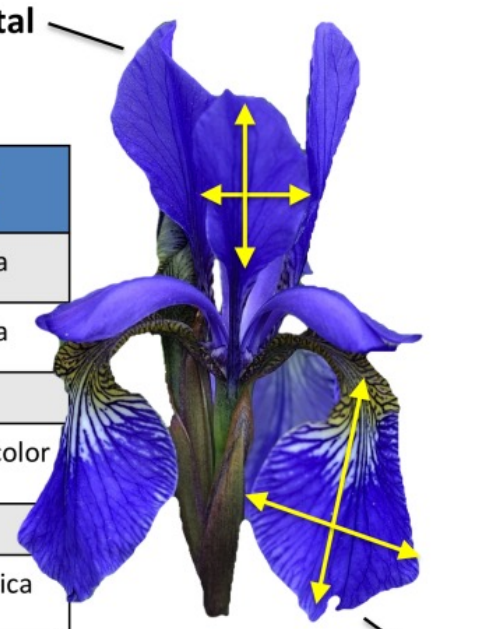


Example: Iris dataset

- 50 samples from each of three class of *Iris (versicolor, setosa, virginica)*
- Four features: sepal length, sepal width, petal length, petal width

Samples
(instances, observations)

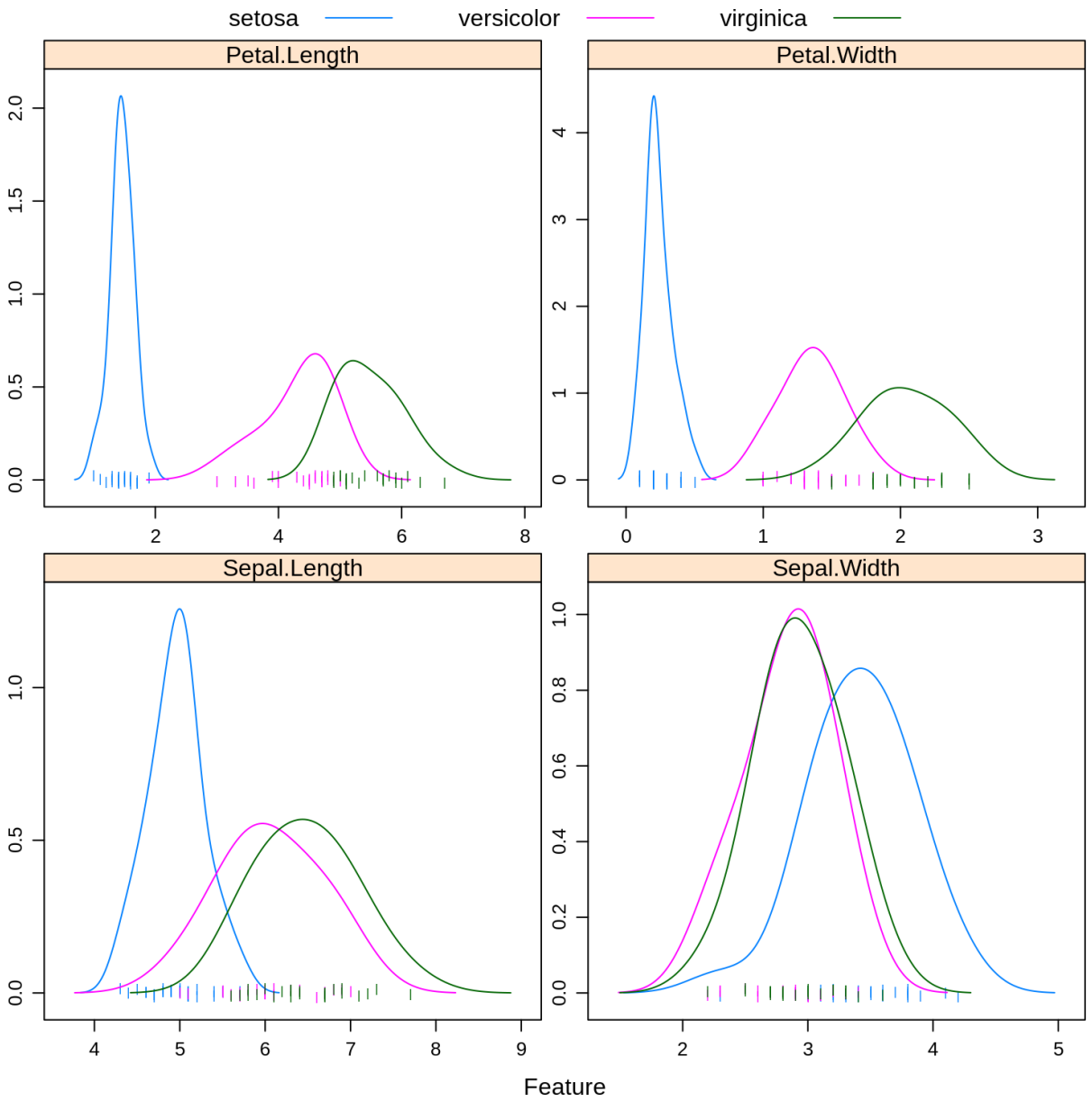
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

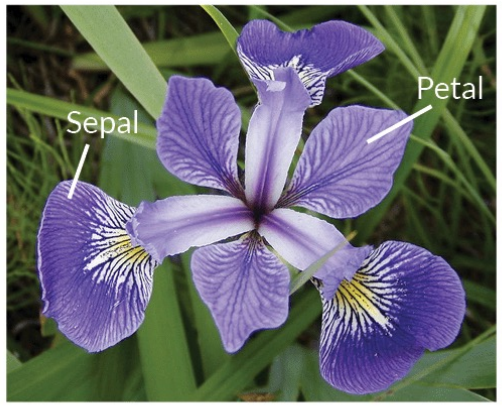


Class labels
(targets)

Features
(attributes, measurements, dimensions)

Distribution of features





Iris Versicolor

Iris Setosa

Iris Virginica

Generative model: Linear discriminant analysis

- Model $\Pr[X = x \mid Y = k]$

$$X = \begin{bmatrix} sepal\ length \\ sepal\ width \\ petal\ length \\ petal\ width \end{bmatrix}$$

$$Y \in \{versicolor, setosa, virginica\}$$

by a **multivariate normal distribution** $N(\mu_k, \Sigma)$ with mean μ_k, covariance matrix Σ

- Model $\Pr[X = x \mid Y = k]$

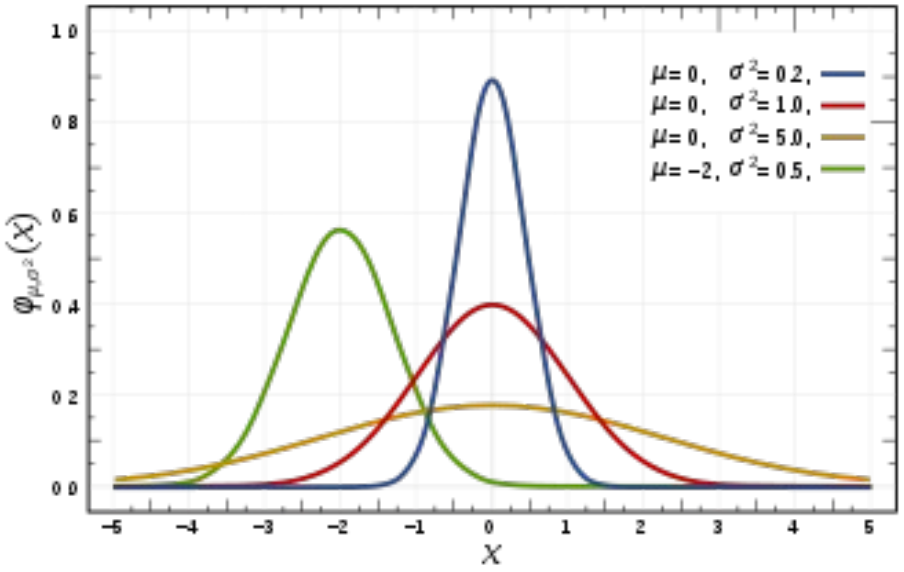
$$X = \begin{bmatrix} sepal\ length \\ sepal\ width \\ petal\ length \\ petal\ width \end{bmatrix}$$

$$Y \in \{versicolor, setosa, virginica\}$$

by a **multivariate normal distribution** $N(\mu_k, \Sigma)$ with mean μ_k, covariance matrix Σ

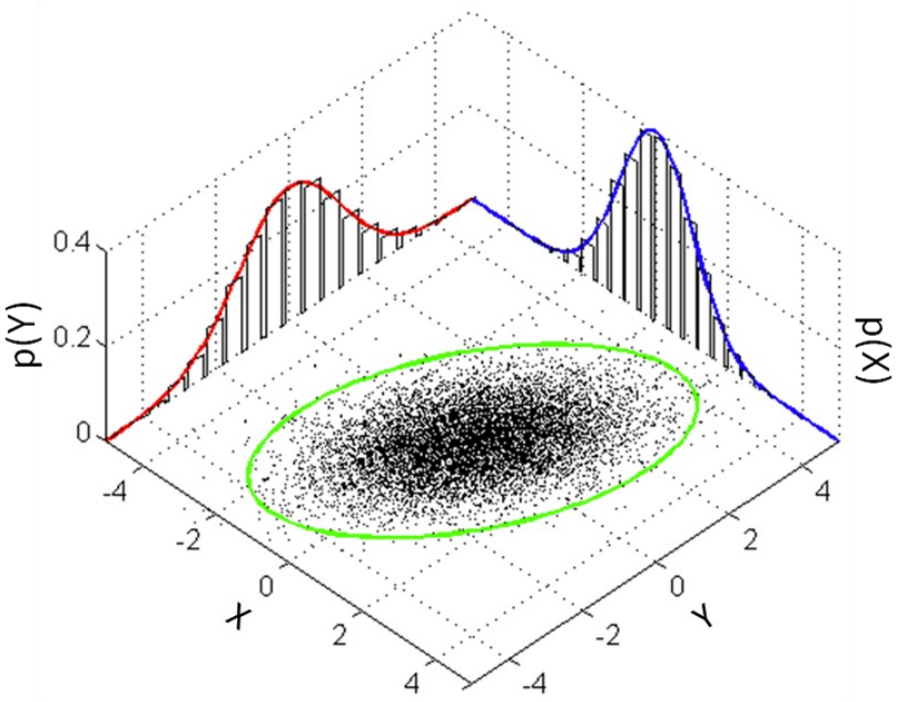
LDA for one-dimensional data

- For the k-th class, model density function as $N(\mu_k, \sigma^2)$
- Density function: $\Pr[X = x|Y = k] = f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$
- Within each class, the features have a center μ_k for every class k and **common variance σ^2**



Multi-dimensional case

- $N(\mu, \Sigma)$ is a multi-dimensional Gaussian with mean μ, covariance Σ: μ is a p-dimensional vector, covariance is a $p \times p$ matrix: $\Sigma = E[xx^\top]$
- Illustration of a two-dimensional multivariate normal distribution
- Two dimensions: blue and red
- Projection to every dimension is still a Gaussian
- Centered at zero



NORTHEASTERN UNIVERSITY
LVX VERITAS VIRTVS
18 98

LDA for multi-dimensional data

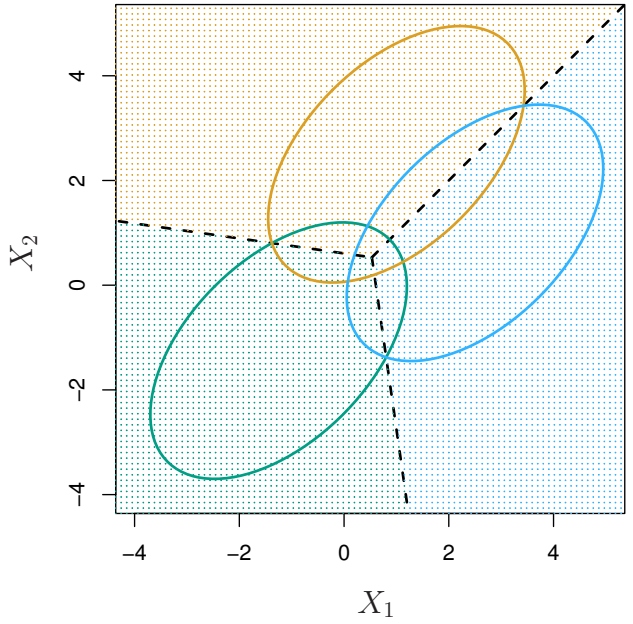
- For the k-th class, model density function as $N(\mu_k, \Sigma)$
- Density function

$$\Pr[X = x|Y = k] = f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \Sigma (x-\mu_k)\right)$$

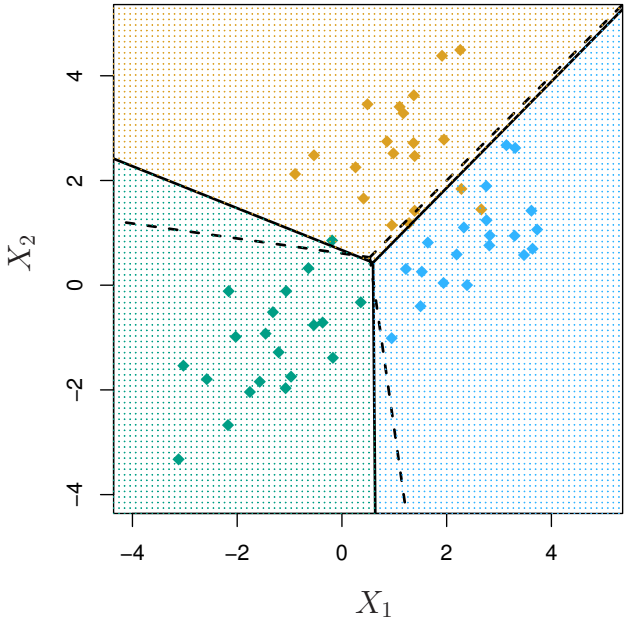
- Within each class, the features have a center μ_k for every class k and **common variance σ^2**

Example

- Example with a two-dimensional synthetic dataset



Dash lines: Bayes decision boundaries



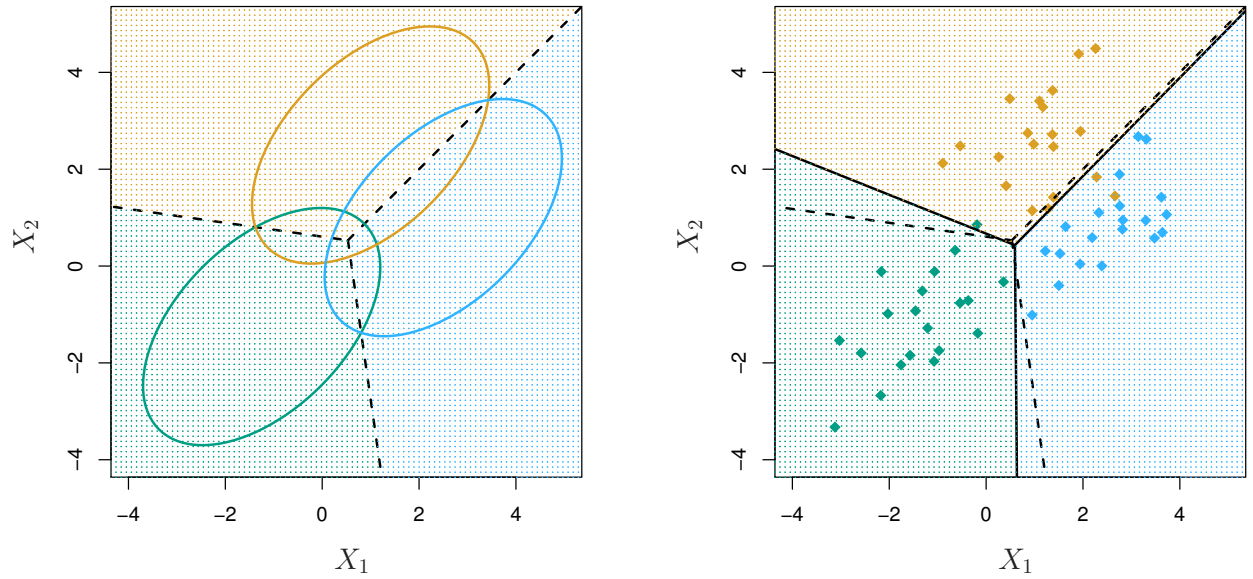
Solid lines: LDA decision boundaries (they are linear)

Estimating the center

How does this work?

1. Estimate the center of each class μ_k:

$$\hat{\mu}_k = \frac{1}{\#\{i: y_i = k\}} \sum_{i:\, y_i = k} x_i$$

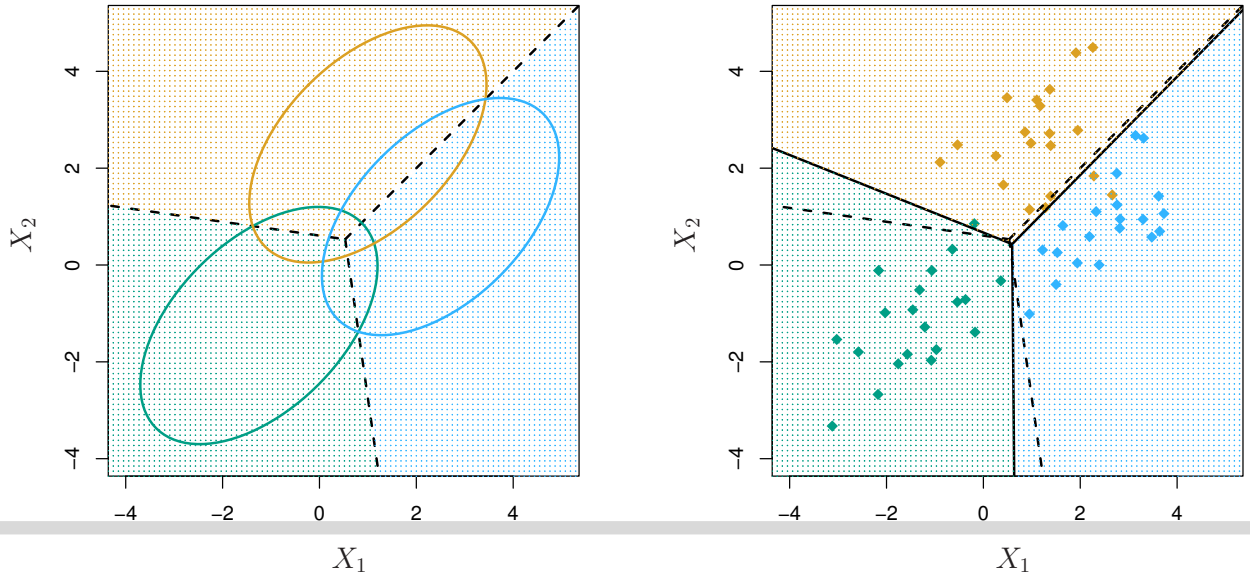


Estimating the covariance

How does this work?

2. Estimate the common covariance matrix Σ

- One-dimensional data: $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \hat{\mu}_k)^2$
- Multi-dimensional data: Compute the vectors of deviations $(x_1 - \hat{\mu}_{y_1}), (x_2 - \hat{\mu}_{y_2}), \cdots, (x_n - \hat{\mu}_{y_n})$ and their covariance

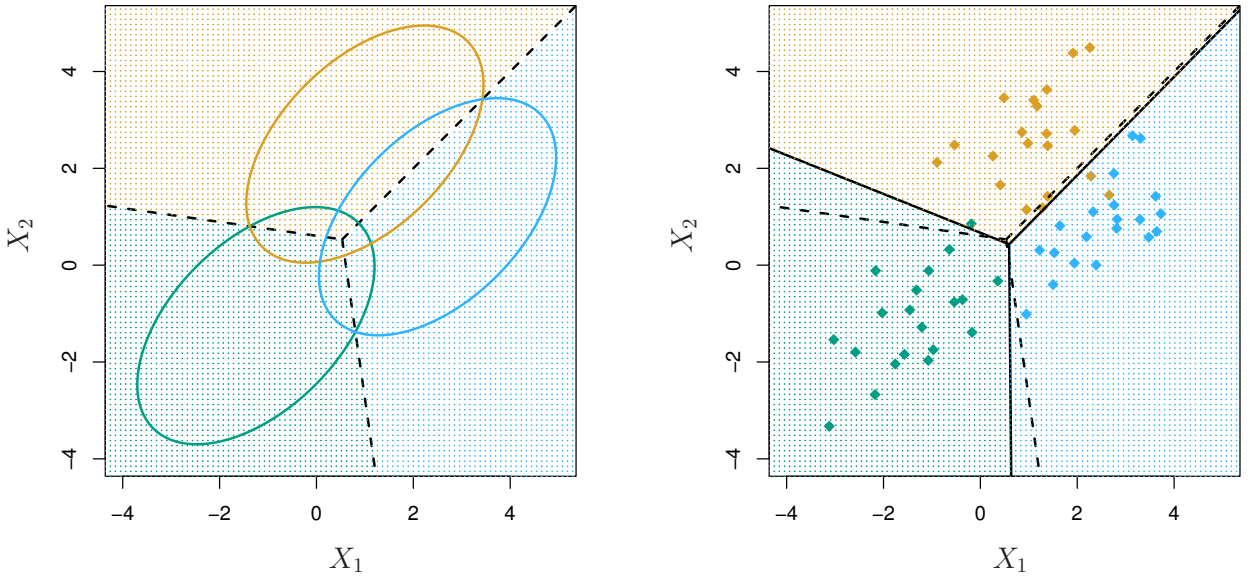


Estimating the prior

How does this work?

3. Estimated by the fraction of training samples of class k: $\Pr[Y = k] = \hat{\pi}_k$

$\hat{\pi}_k = \frac{\#\{i: y_i = k\}}{n}$: Fraction of training samples of class k



Prediction

- Recall: $\Pr(Y = k|X = x)$ is probability of x having label k
- LDA predicts the label with highest probability
- We use Bayes rule

$$\Pr[Y = k|X = x] = \frac{Pr(Y = k, X = x)}{Pr(X = x)} = \frac{Pr(X = x|Y = k) \cdot Pr(Y = k)}{\sum_{i=1}^{K} Pr(X = x|Y = i) \cdot Pr(Y = i)}$$

Recap

- Mixture of Gaussian model: A prior probability for each cluster, with one Gaussian distribution for the density of every cluster
- The clusters have a separate center and a common covariance matrix
- Parameter estimation involves estimating the mean, covariance, and prior of the mixture of Gaussian model
- We have shown that we can estimate each with simple statistics from data

Announcements

- TA office hours
  - Monday, 1 PM – 1:45 PM, WVH 208 (or via Zoom)
  - Wednesday, 1 PM – 1:45 PM, WVH 208 (or via Zoom)
  - Friday, 12:30 PM – 1:30 PM, 22nd floor, 177 Huntington Ave (or via Zoom)

- Join piazza! Access code: 8128pzbevas. Signup link: https://piazza.com/northeastern/fall2024/ds522020725202510