Supervised Machine Learning and Learning Theory

Lecture 5: Classification (continued)

September 20, 2024

Example: Iris dataset

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

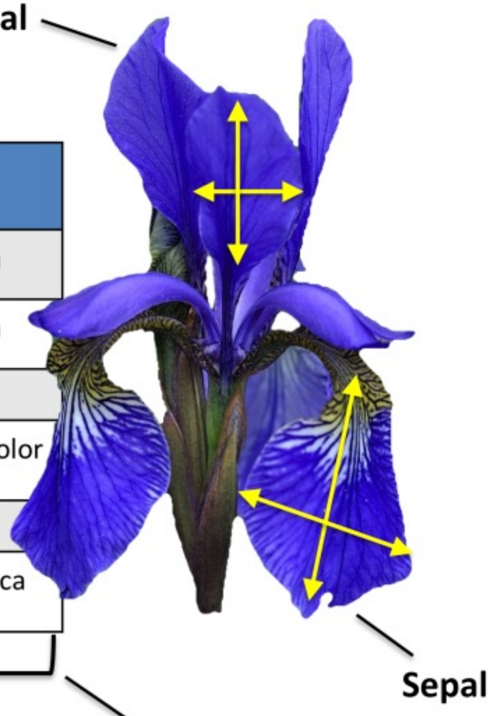
Iris Versicolor **Iris Setosa** **Iris Virginica**

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)

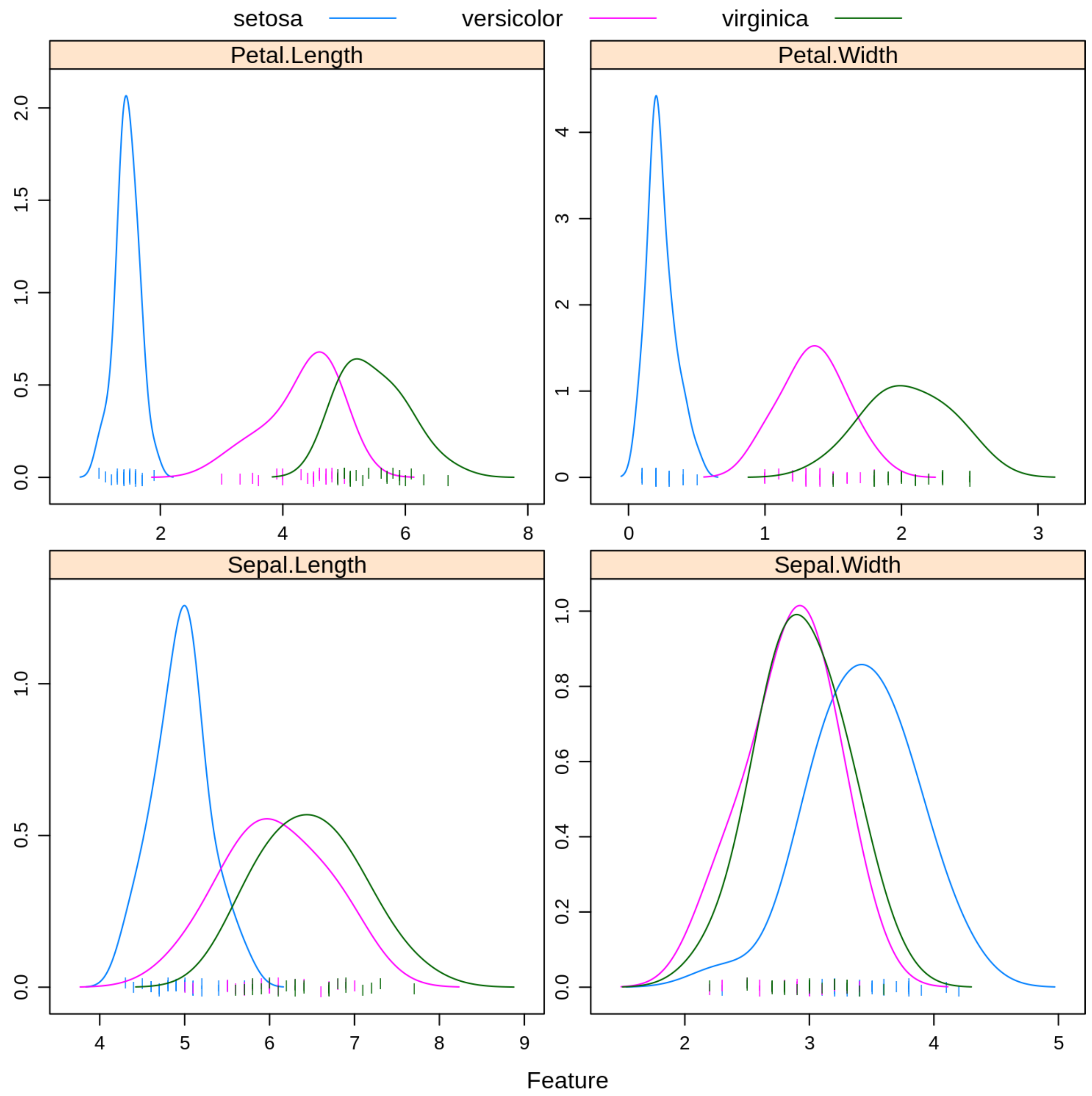
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Distribution of features



Iris Versicolor **Iris Setosa** **Iris Virginica**

Fit a mixture of Gaussians to each feature

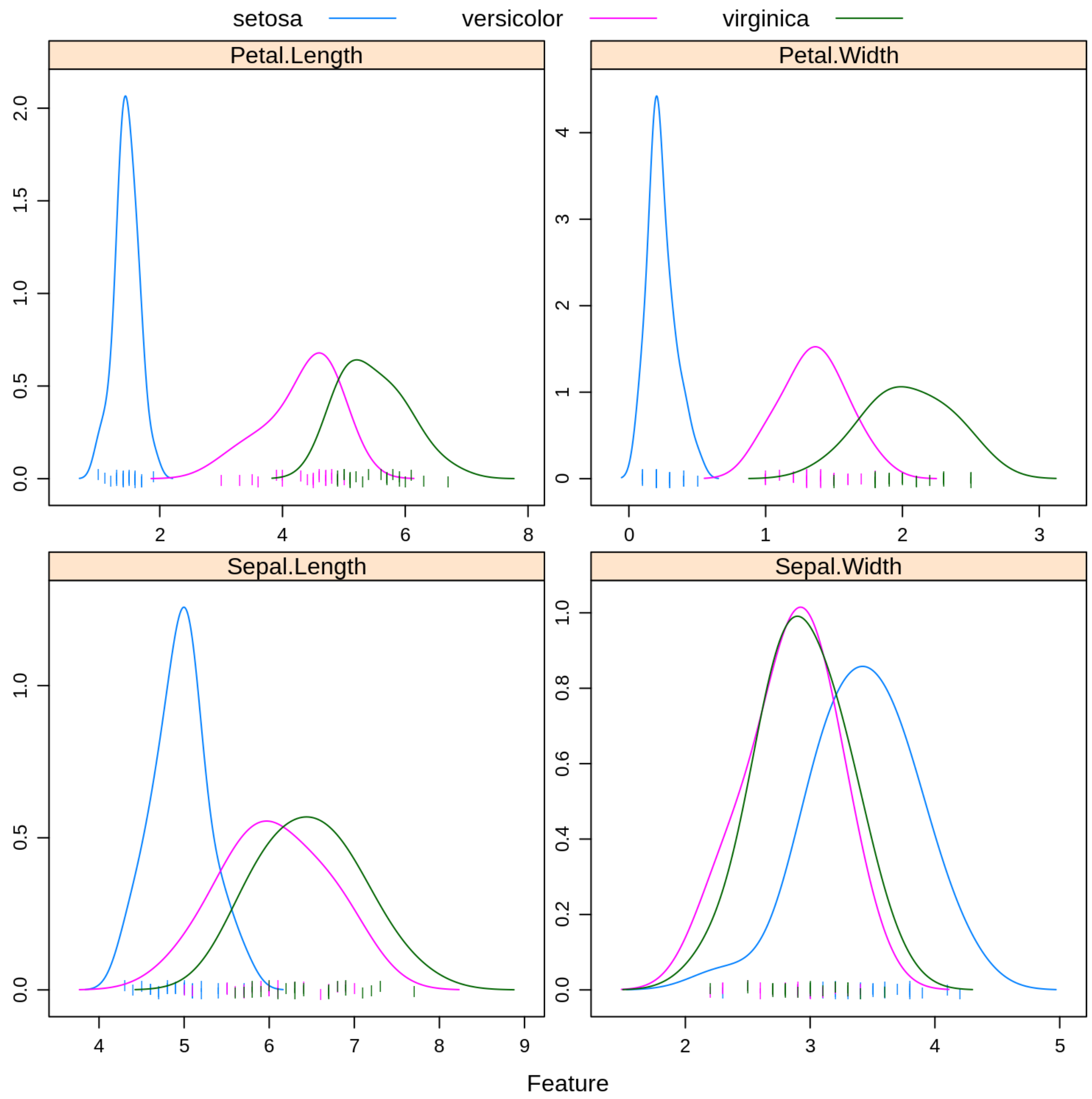
- Model $\Pr[X = x \mid Y = k]$

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

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

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

- $N(\mu_k, \Sigma)$ denotes a Gaussian distribution



Prediction rule: Use the Bayes rule

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

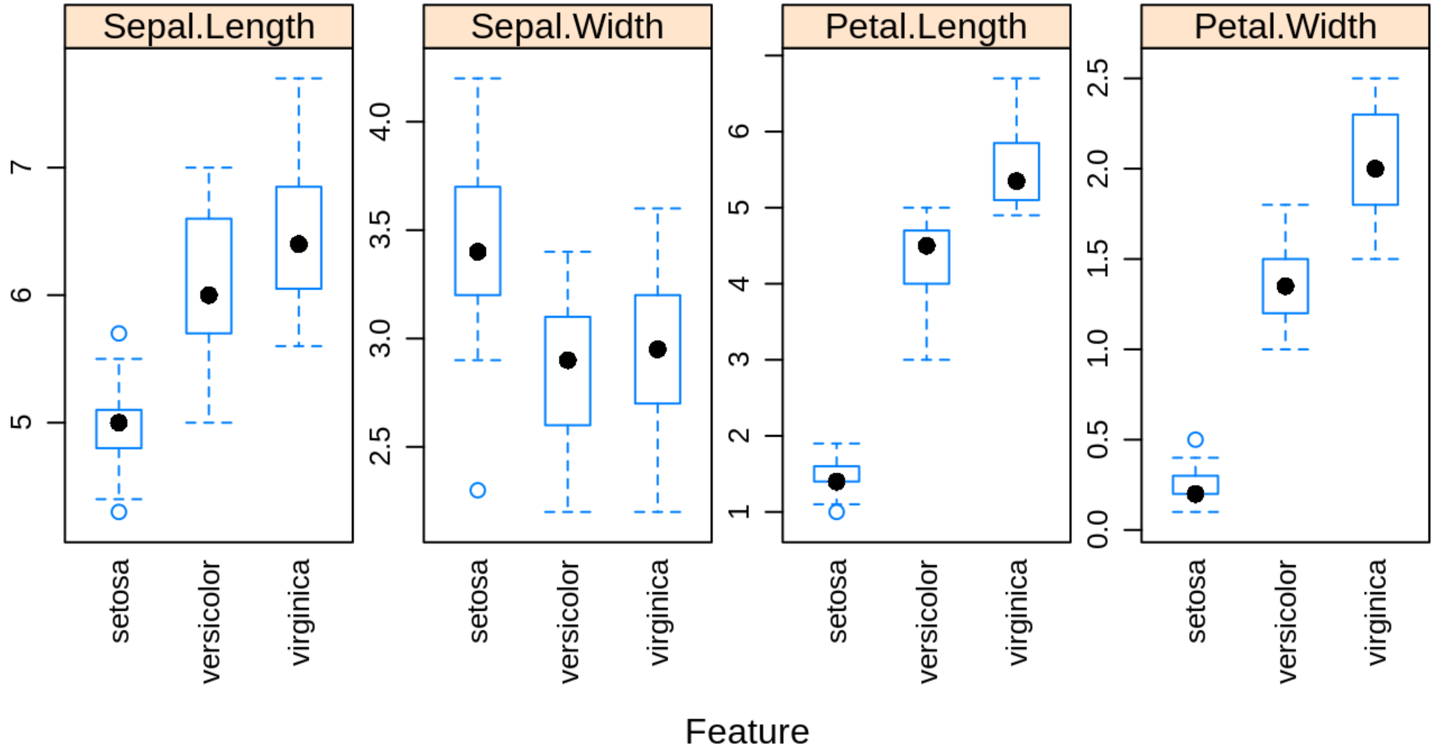
$$\Pr[Y = k|X = x] = \frac{Pr(Y = k, X = x)}{Pr(X = x)} = \frac{Pr(X = x|Y = k) \cdot Pr(Y = k)}{\sum_{i=1}^{K} Pr(X = x|Y = i) \cdot Pr(Y = i)}$$

  - Examples of conditional probability: Conditioned on the weather is rainy, the chance that driving time is extended would higher than if the weather is sunny

Illustrating μ_k in iris dataset

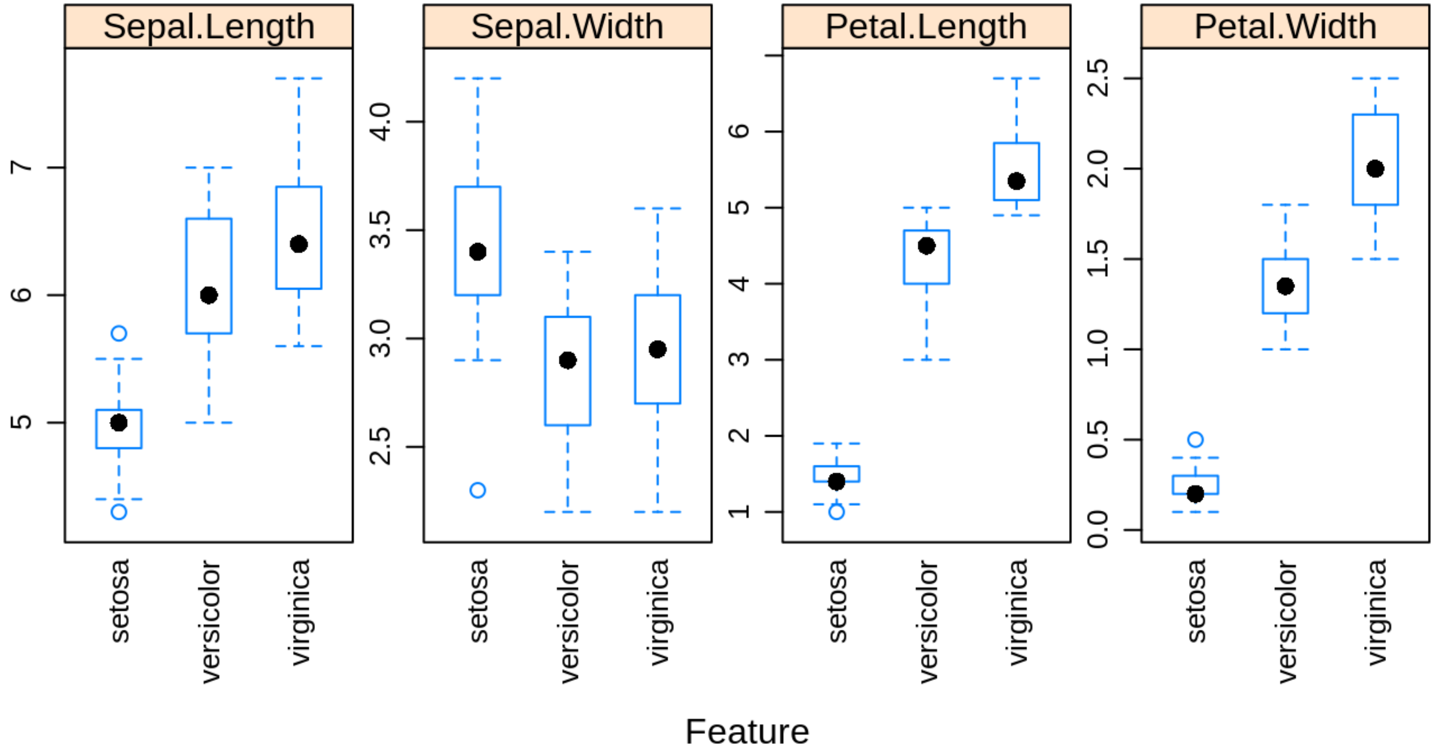
$$\mu_{setosa} = \begin{bmatrix} \overline{setosa\ sepal\ length} \\ \overline{setosa\ sepal\ width} \\ \overline{setosa\ petal\ length} \\ \overline{setosa\ petal\ width} \end{bmatrix}$$

- Bar represents average value
- Black dots of setosa in the box plots



Illustrating Σ in iris dataset

- Σ is the same for $versicolor, setosa, virginica$
  - Diagonal entries equal to variance of each feature for all classes: Proportional to the width of the box plots
  - Off-diagonal entries equal to covariance between two features for all classes (somewhat like correlation coefficients)



- **Question:** What if Σ should be different for different class?

Linear decision boundaries

- Prior probability: $\Pr[Y = k] = \pi_k$
- Density function: $\Pr[X = x|Y = k]$ is multivariate normal $N(\mu_k, \Sigma)$, where μ_k: mean for category k, Σ: covariance matrix

- The density function for the k-th class follows the multivariate normal distribution:

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \cdot \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}$$

Why LDA has linear decision boundaries

- According to Bayes rule: $Pr(Y = k|X = x) = \frac{Pr(X=x|Y=k) \cdot Pr(Y=k)}{\sum_{i=1}^{K} Pr(X=x|Y=i) \cdot Pr(Y=i)}$

- Take the log on both sides: $\log \Pr[Y = k|X = x] = \log[\Pr[X = x|Y = k]] + \log[\Pr[Y = k]] - \log\left[\sum_{i=1}^{K} \Pr[X = x|Y = i] \cdot \Pr[Y = i]\right]$

- Decision boundary corresponds to $\log \Pr[Y = k|X = x] = \log \Pr[Y = l|X = x]$ between class k and class j

- The third term cancels out. This leaves us with:

$$\log[\Pr[X = x|Y = k]] + \log[\Pr[Y = k]] = \log[\Pr[X = x|Y = l]] + \log[\Pr[Y = l]]$$

- Left-hand side is $-\frac{1}{2}(x - \mu_k)^{\top}\Sigma^{-1}(x - \mu_k) + \log \pi_k - \log\left((2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{p}{2}}\right)$

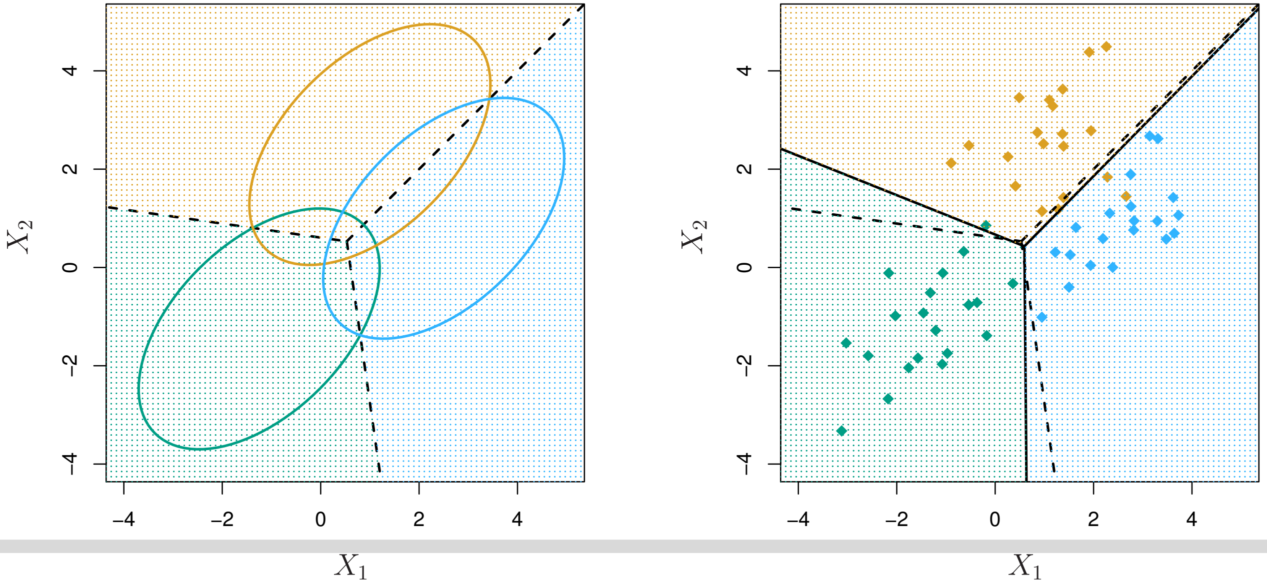
Why LDA has linear decision boundaries

- Decision boundary given by

$$\log \pi_k - \frac{1}{2}\mu_k{}^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k = \log \pi_l - \frac{1}{2}\mu_l{}^T\Sigma^{-1}\mu_l + x^T\Sigma^{-1}\mu_l$$

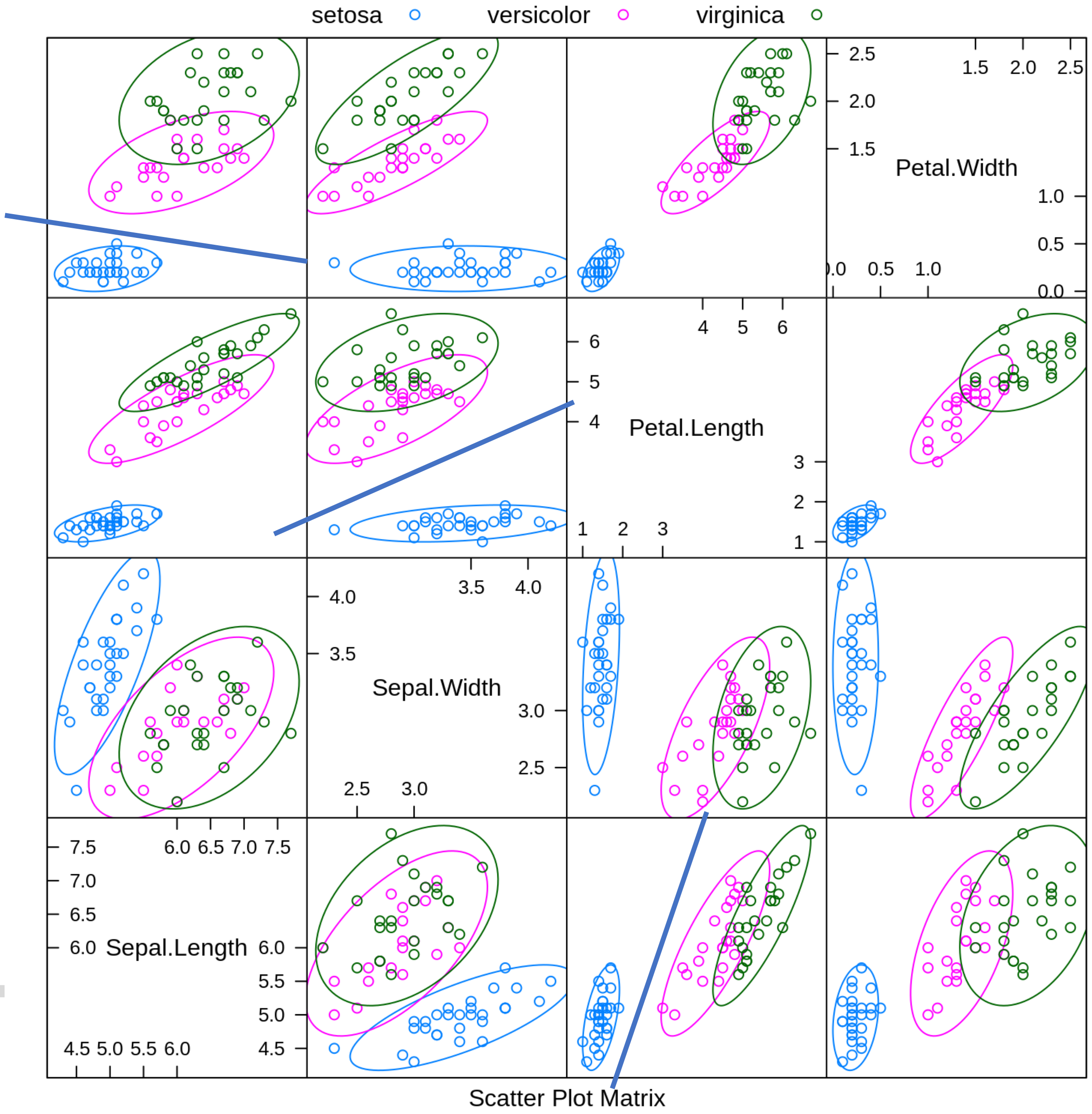
- This is linear in x:

$$x^\top\Sigma^{-1}(\mu_k - \mu_l) = \log \pi_l - \log \pi_k + \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_l^\top\Sigma^{-1}\mu_l$$



LDA decision boundaries for iris dataset

- Illustration of linear boundaries for separate three classes



Quadratic discriminant analysis

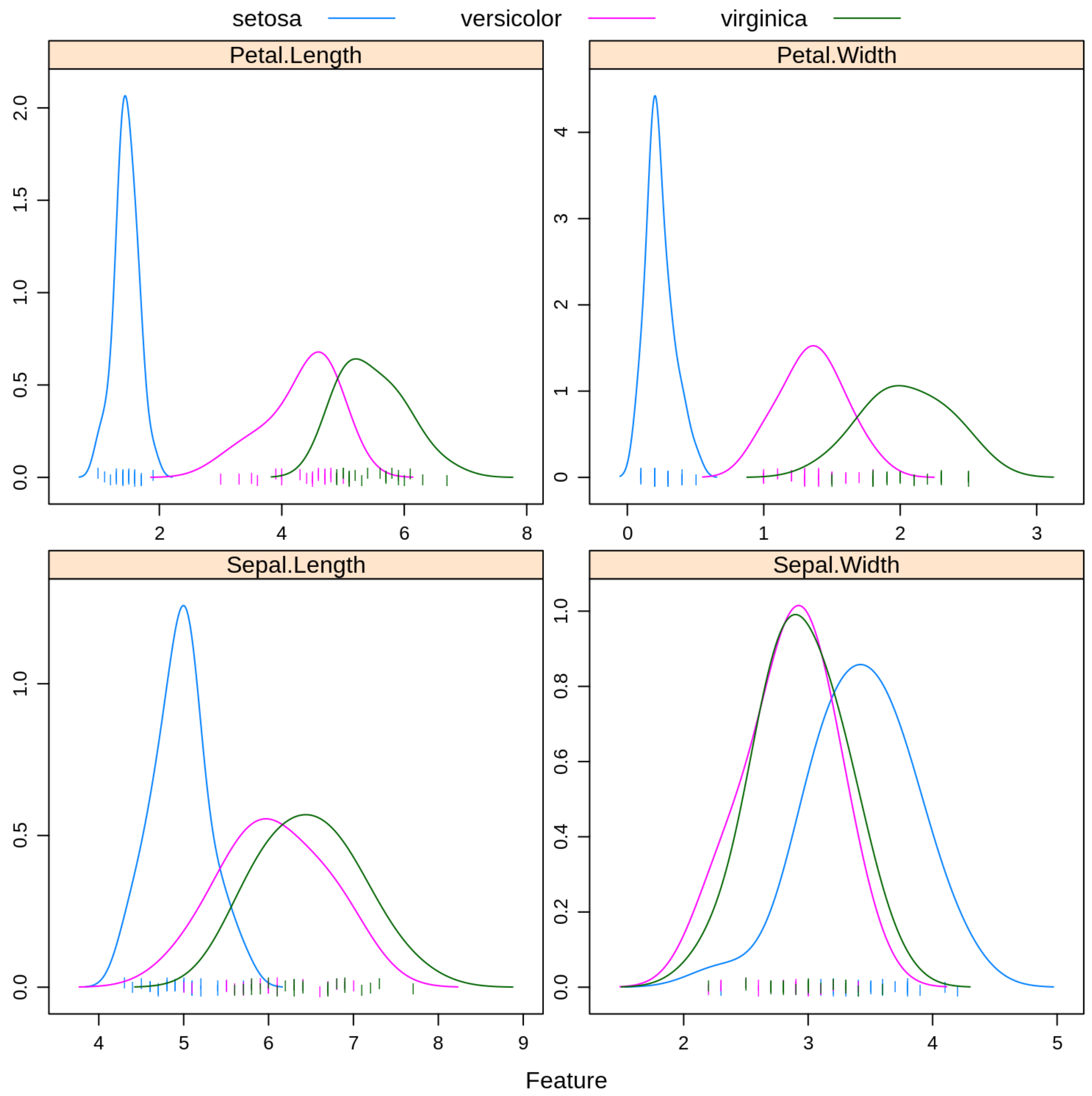
- Model $P(X = x \mid Y = k)$

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

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

by a multivariate normal distribution $N(\mu_k, \Sigma_k)$ with mean μ_k, covariance matrix Σ_k

- Using a different covariance matrix for each class

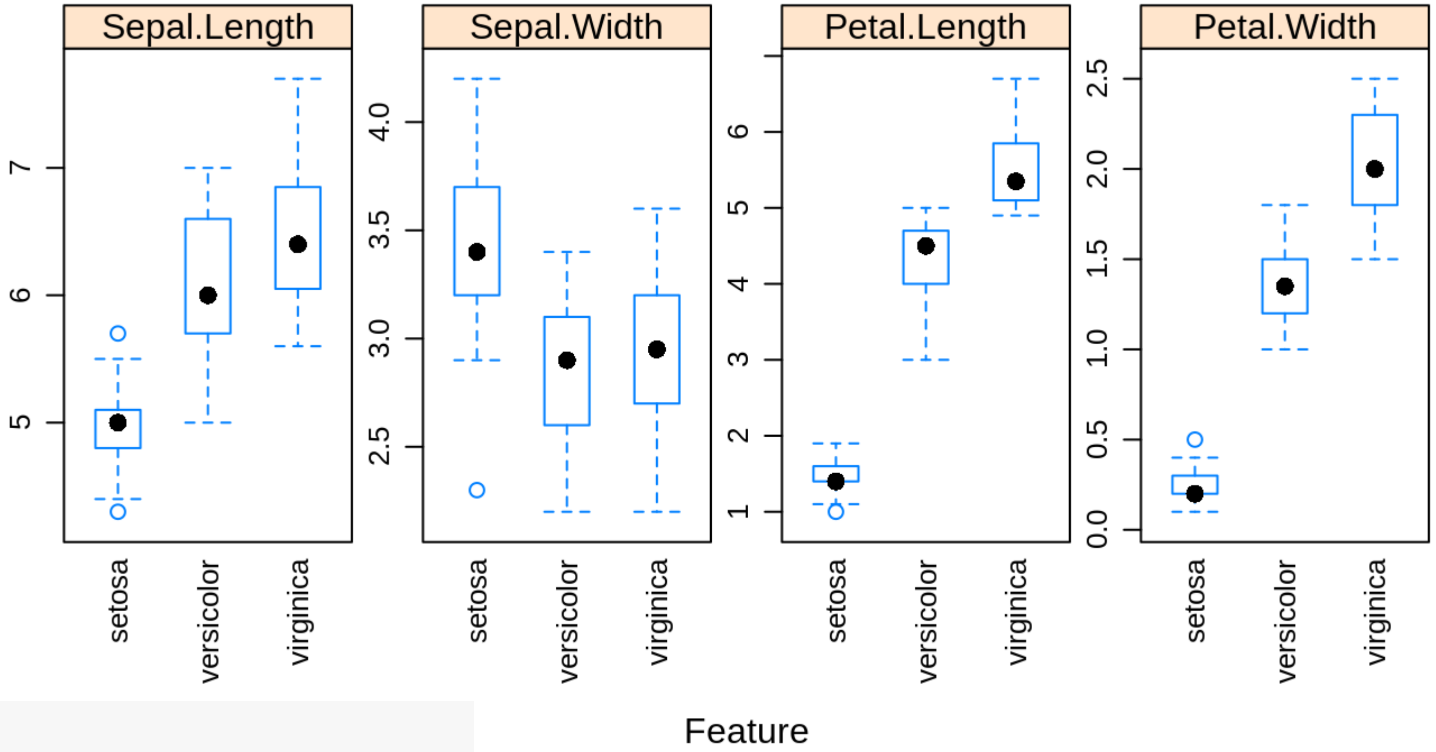


QDA: Estimating the center μ_k

- Estimate the center of each class μ_k:

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

where $n_k = \#\{i: y_i = k\}$




```
## Group means:
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa         4.958621    3.420690     1.458621    0.237931
## versicolor     6.063636    2.845455     4.318182    1.354545
## virginica      6.479167    2.937500     5.479167    2.045833
```

QDA: Estimating the covariance Σ_k

- Estimate the covariance Σ_k

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k) \cdot (x_i - \hat{\mu}_k)^\top$$

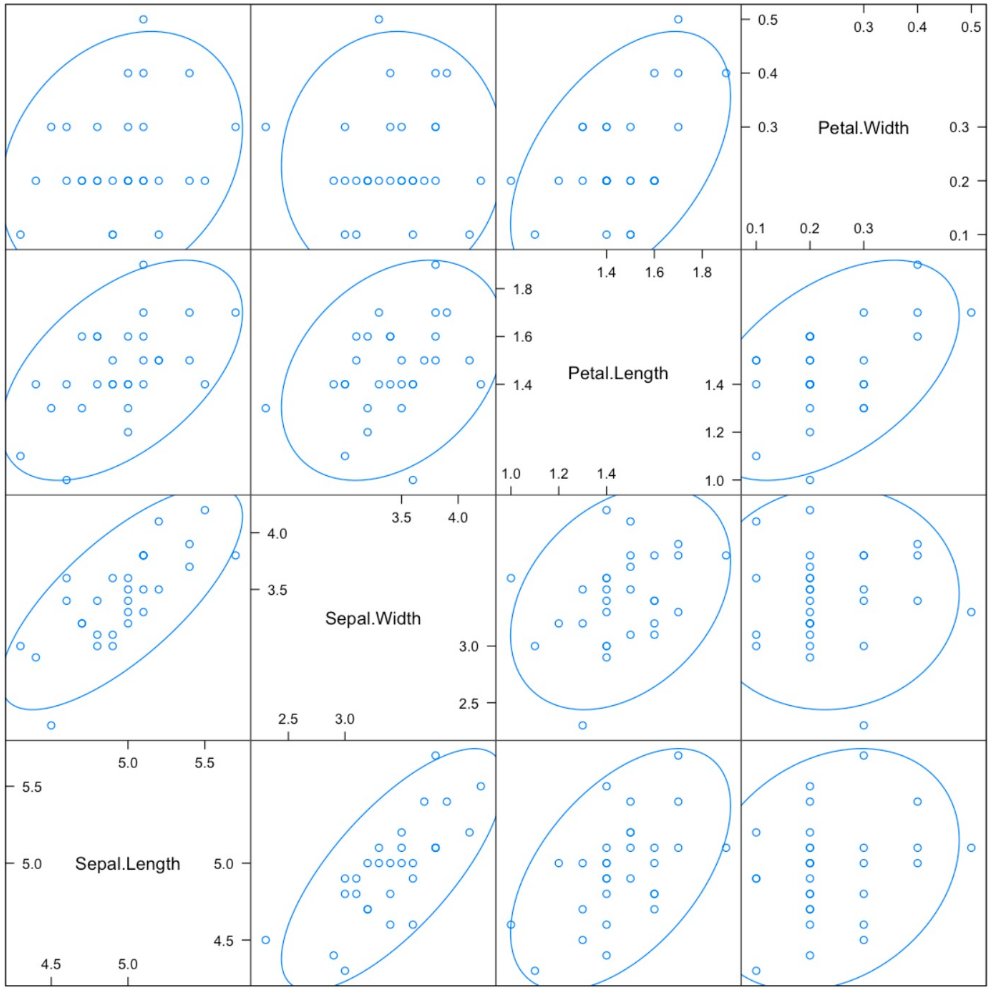
where $n_k = \#\{i: y_i = k\}$

- Example: Σ_{setosa}

```
iris_trn_setosa <- iris_trn[iris_trn$Species == "setosa",]
cov(iris_trn_setosa[,c(1:4)])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length  0.103226601 0.095172414  0.031798030 0.007697044
## Sepal.Width   0.095172414 0.160985222  0.025172414 0.001687192
## Petal.Length  0.031798030 0.025172414  0.035369458 0.009125616
## Petal.Width   0.007697044 0.001687192  0.009125616 0.009581281
```

setosa



Scatter Plot Matrix

Summary of QDA

- For each class k, we model $Pr(X = x|Y = k) = f_k(x)$ as a multivariate normal distribution $N(\mu_k, \Sigma_k)$ with mean μ_k and a different covariance matrix Σ_k

- We estimate $Pr(X = x|Y = k)$ as $N(\hat{\mu}_k, \hat{\Sigma}_k)$ and $Pr(Y = k) = \hat{\pi}_k$

- We apply Bayes rule to obtain $Pr(Y = k \mid X = x)$

$$Pr(Y = k \mid X = x) = \frac{Pr\,(Y = k, X = x)}{Pr\,(X = x)} = \frac{Pr(X = x \mid Y = k)Pr\,(Y = k)}{\sum_j Pr(X = x \mid Y = j)Pr\,(Y = j)}$$

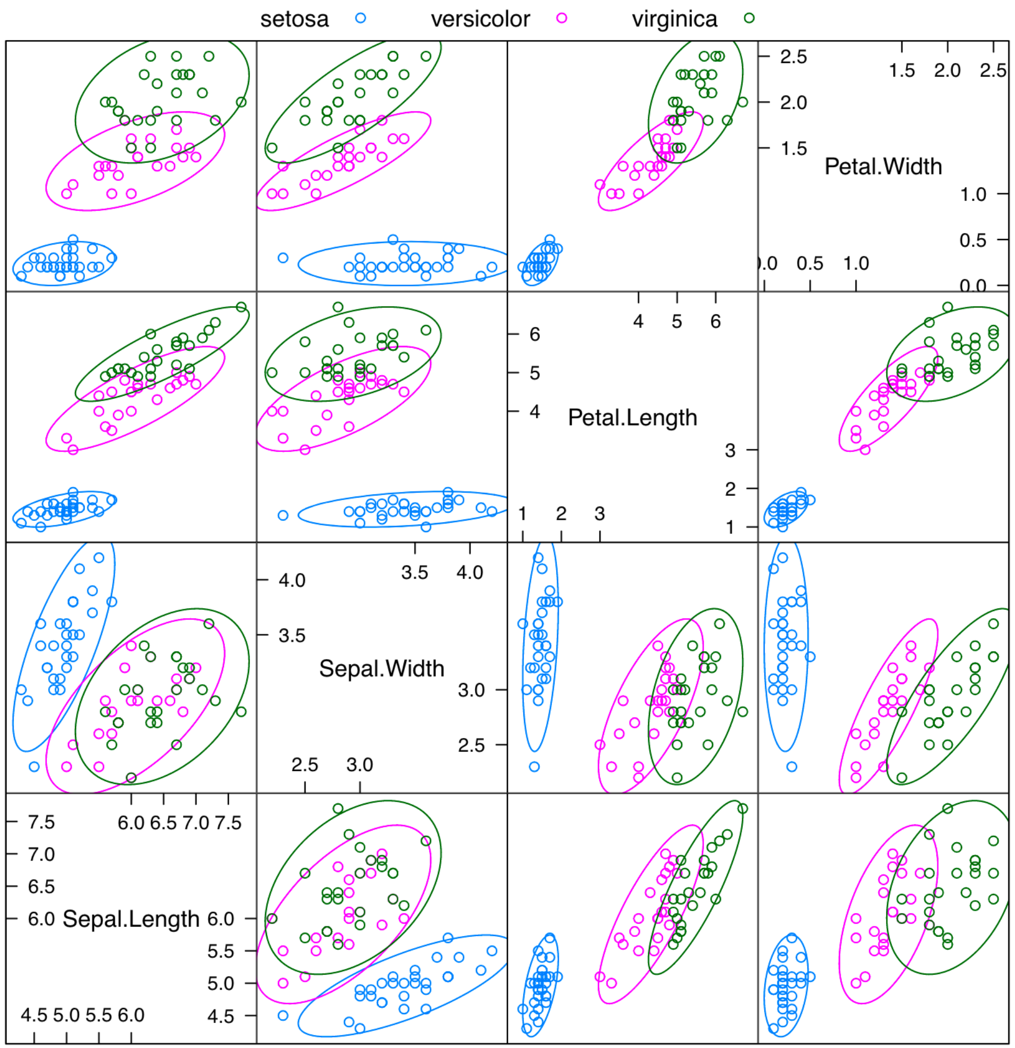
Covariance in LDA vs. in QDA

- In LDA, the covariance can also be estimated directly as follows:

$$\hat{\Sigma} = \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\Sigma}_k$$

where $n_k = \#\{i: y_i = k\}$

- $\hat{\Sigma} = \frac{n_{setosa}-1}{n-3} \cdot \hat{\Sigma}_{setosa} + \frac{n_{versicolor}-1}{n-3} \cdot \hat{\Sigma}_{versicolor} + \frac{n_{virginica}-1}{n-3} \cdot \hat{\Sigma}_{virginica}$



Scatter Plot Matrix

Decision boundaries for QDA are quadratic

- For QDA, with some algebra (similar to our calculation for LDA), let

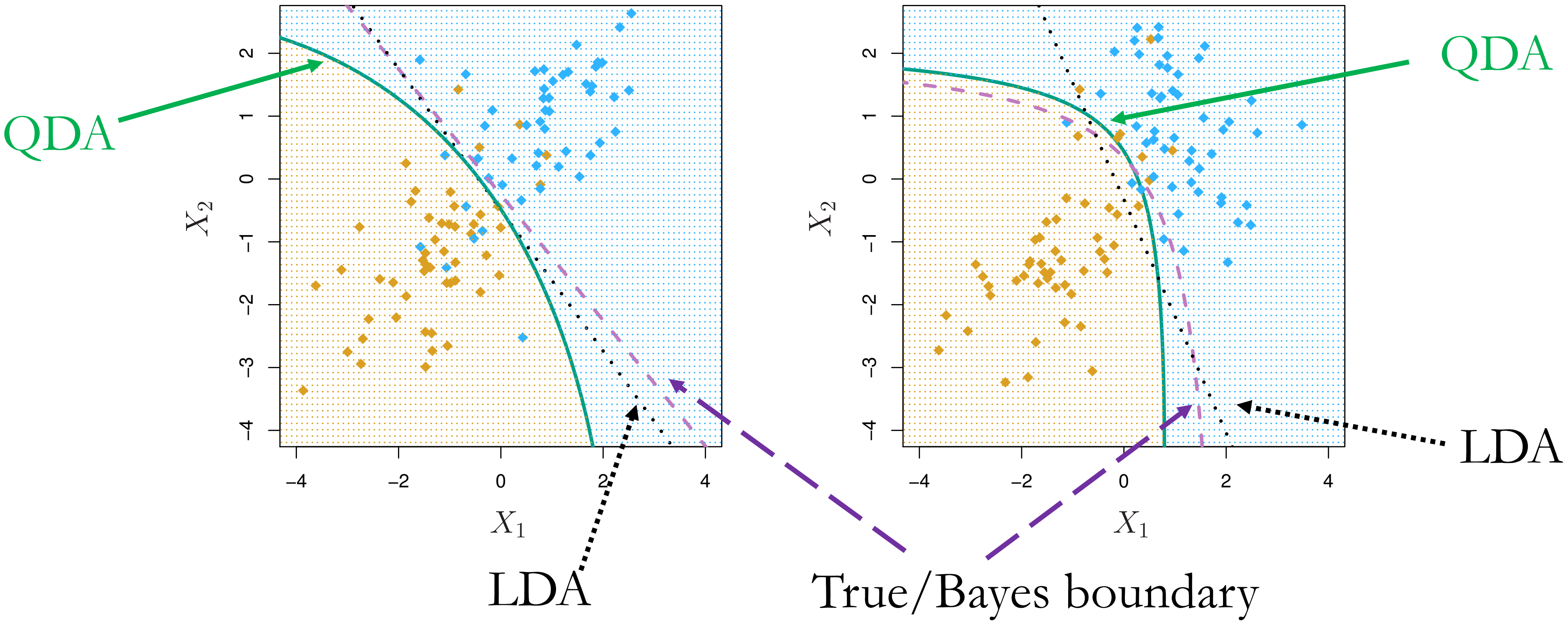
$$\log Pr(Y = k|X = x) = C + \hat{\delta}_k(x)$$

where $\delta_k(x) = \log \pi_k - \frac{1}{2}\mu_k{}^T\Sigma_k^{-1}\mu_k + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}x^T\Sigma_k^{-1}x - \frac{1}{2}\log|\Sigma_k|$ and C is a constant

- $\hat{\delta}_k(x)$ is quadratic in x
- Decision boundaries for QDA are quadratic: by setting $\hat{\delta}_k(x) = \hat{\delta}_j(x)$
- For LDA, the quadratic terms would have canceled out

Comparison between LDA and QDA

- QDA requires estimating more model parameters, LDA is less flexible but has a smaller variance



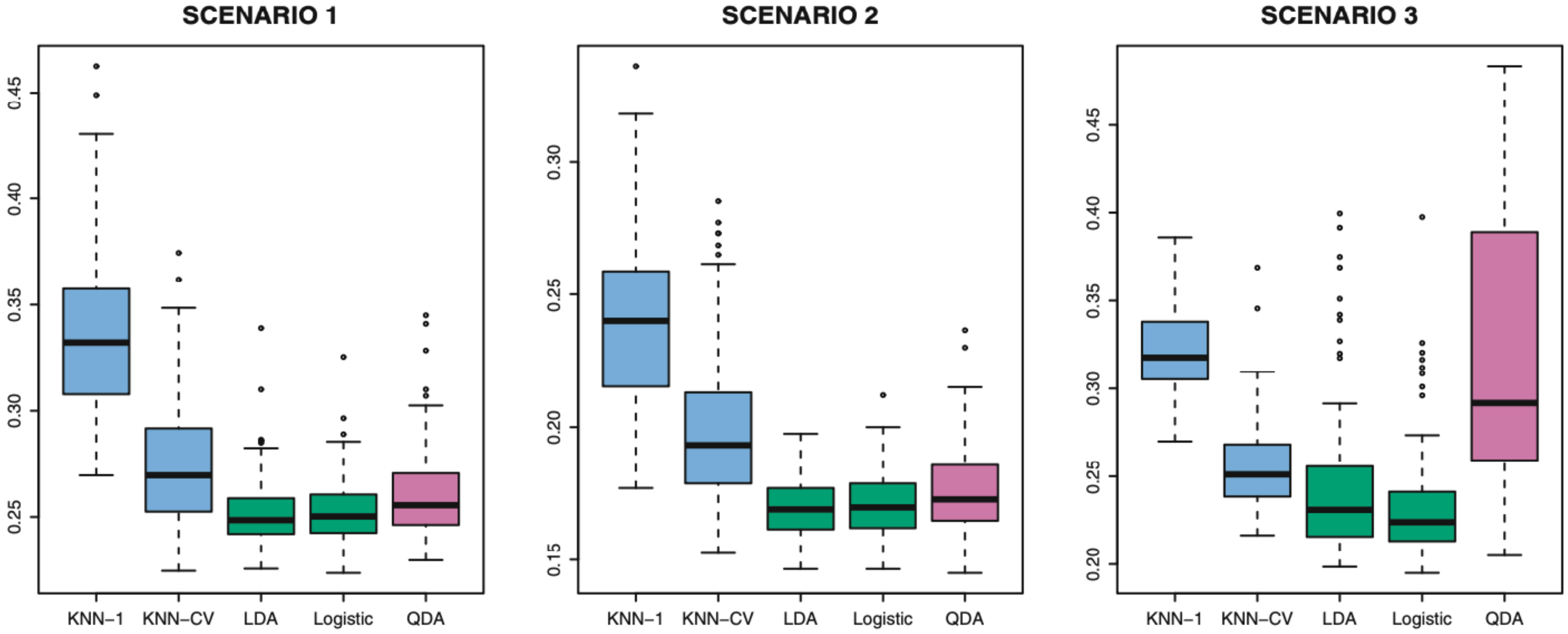
Examples: True decision boundaries are linear

- Data generating process: two predictors X_1 and X_2, two classes in Y

X_1 and X_2 are drawn from uncorrelated Normal distributions with a different mean in each class

Same as Scenario 1, but correlation between X_1 and X_2 is -0.5

X_1 and X_2 are sampled from t-distribution



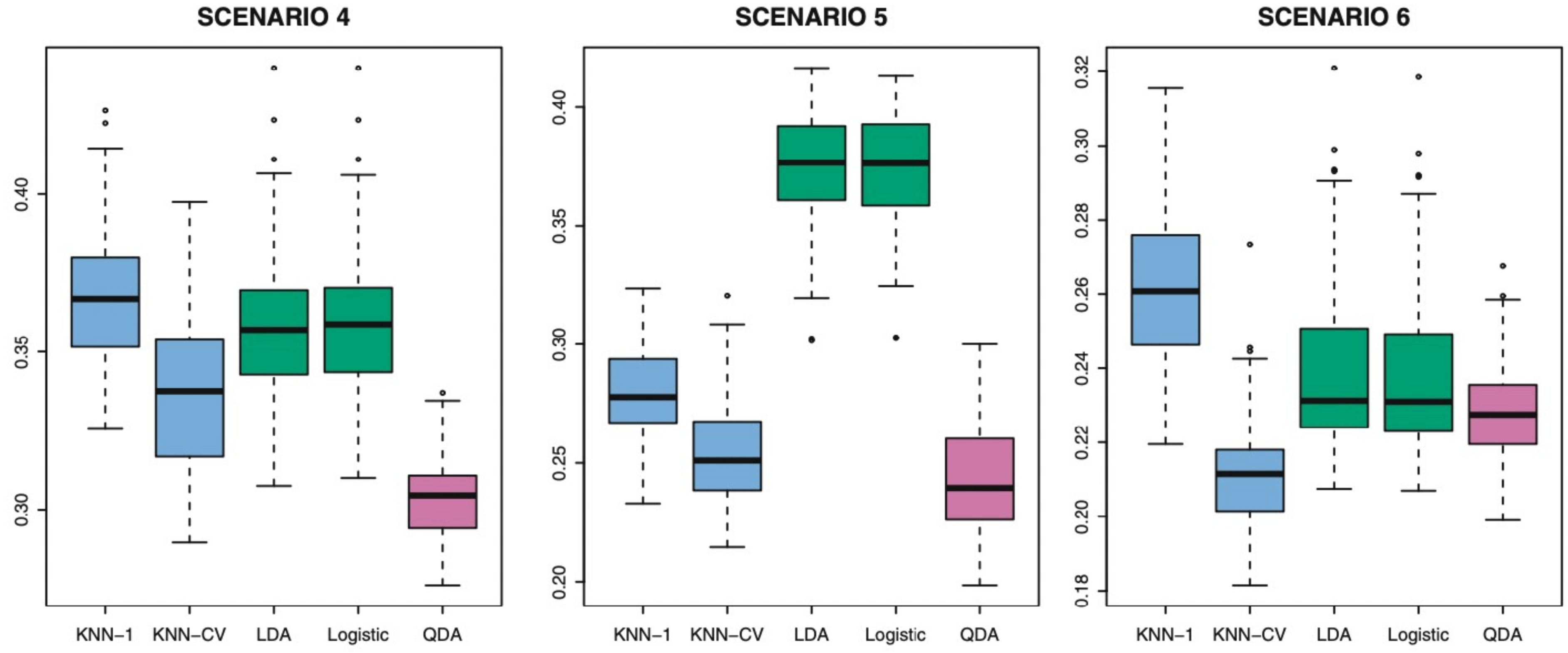
Examples: True decision boundaries are nonlinear

- Data generating process: two predictors X_1 and X_2, two classes in Y

X_1 and X_2 are draw from Normal distributions. First class: correlation between X_1 and X_2 is -0.5. Second class: correlation is 0.5

X_1 and X_2 are drawn from uncorrelated Normal distributions. Y is sampled from logic model using X_1^2, X_2^2 and $X_1 X_2$

Same as Scenario 5. Y is sampled from a more complicated nonlinear function



LDA vs. logistic regression

- Both LDA and logistic regression produce linear decision boundaries

- **Exercise:** Why is the decision boundary of logistic regression linear?
  - Recall logistic regression follows the following log ratio:

$$\log \frac{Pr(Y = 1|X = x)}{Pr(Y = 0|X = x)} = \beta_0 + \beta_1 x$$

  - The decision boundary is the set of x satisfy $Pr(Y = 1|X = x) = Pr(Y = 0|X = x) = 0.5$

$$0 = \log[Pr(Y = 1|X = x)] - \log[Pr(Y = 0|X = x)] = \log \frac{Pr(Y = 1|X = x)}{Pr(Y = 0|X = x)} = \beta_0 + \beta_1 x$$

This is linear in x!

LDA vs. logistic regression

- Estimation approaches are different: **generative** vs. **discriminative**

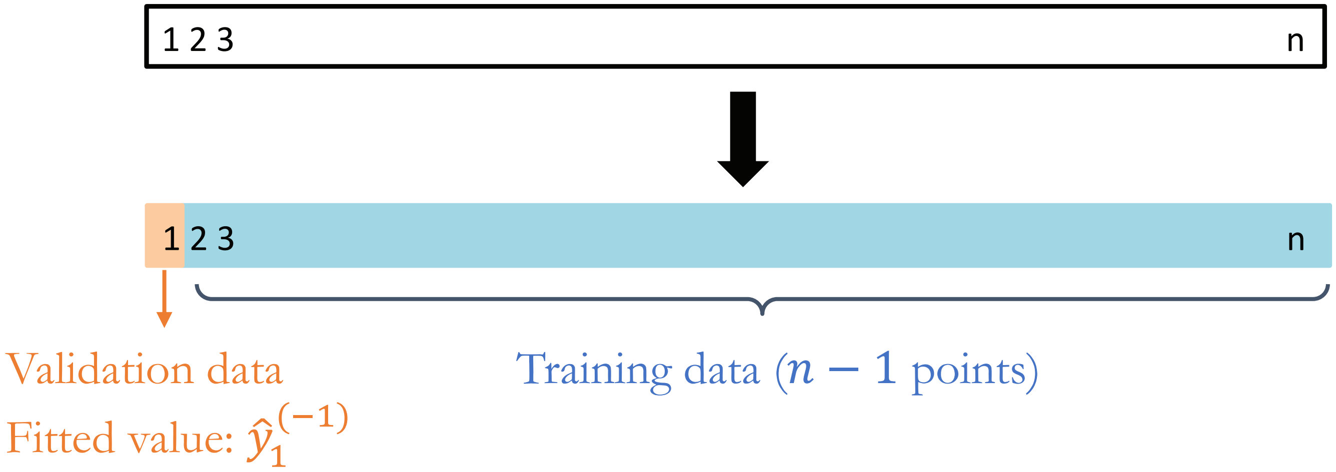
- LDA makes more sense if the underlying data indeed follows a Gaussian distribution (e.g., think of natural data arising in biology)

- Logistic regression is usually more commonly used in practice

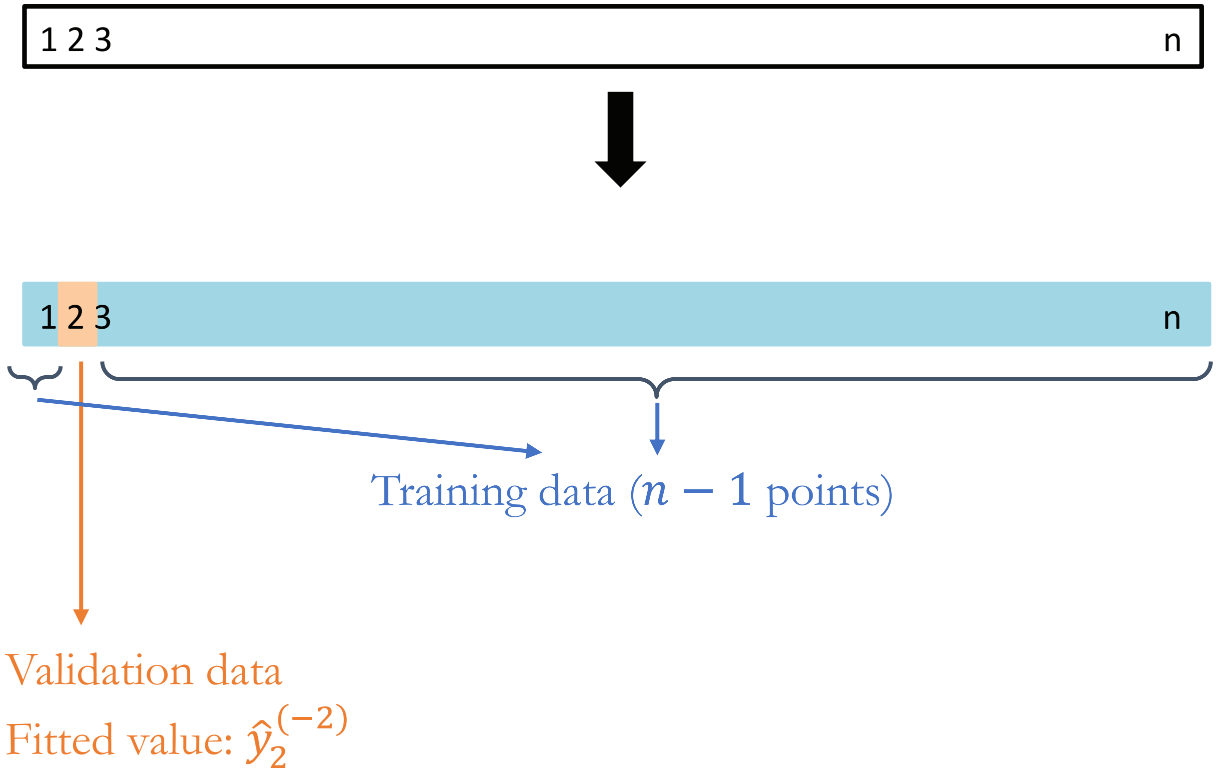
Lecture plan

- Leave-one-out cross-validation
  - For selecting between different models

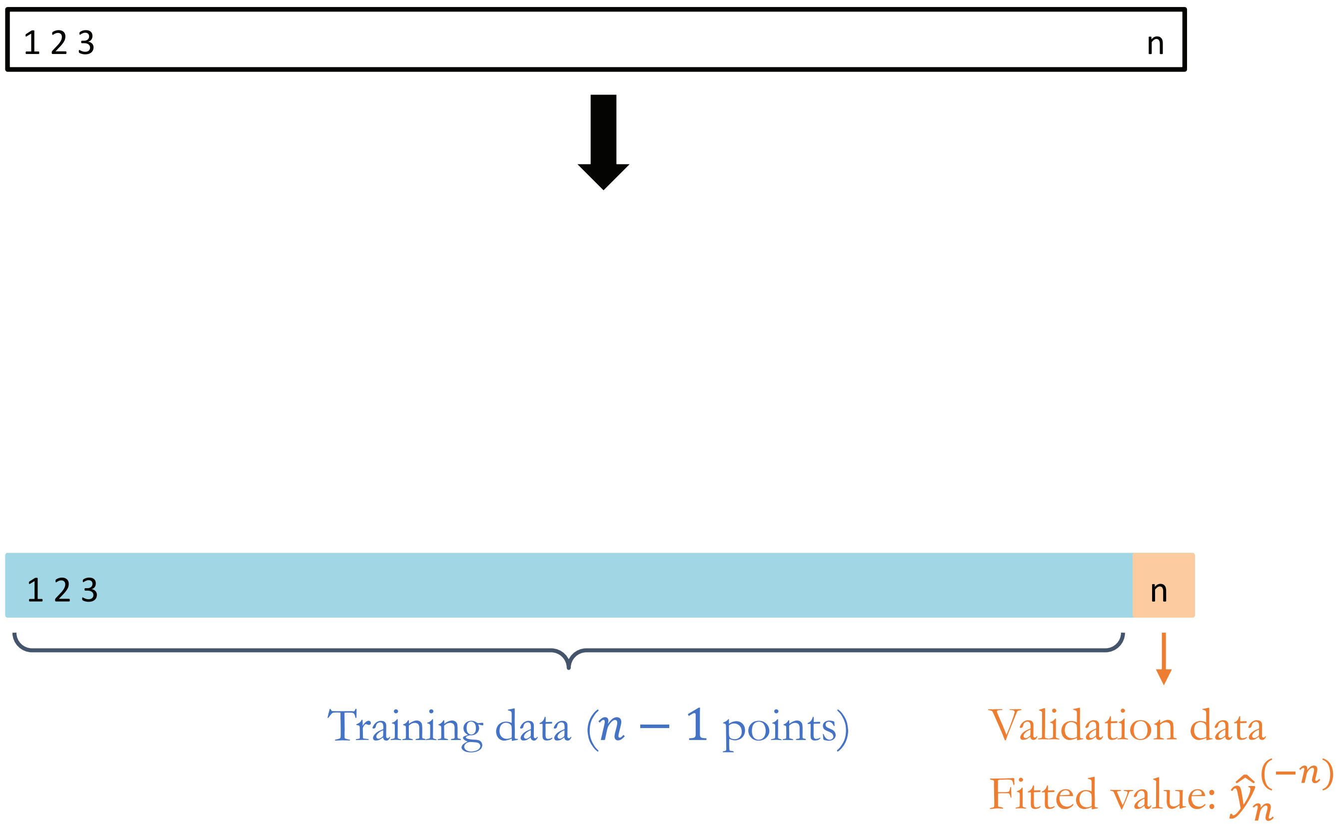
Leave one out cross-validation



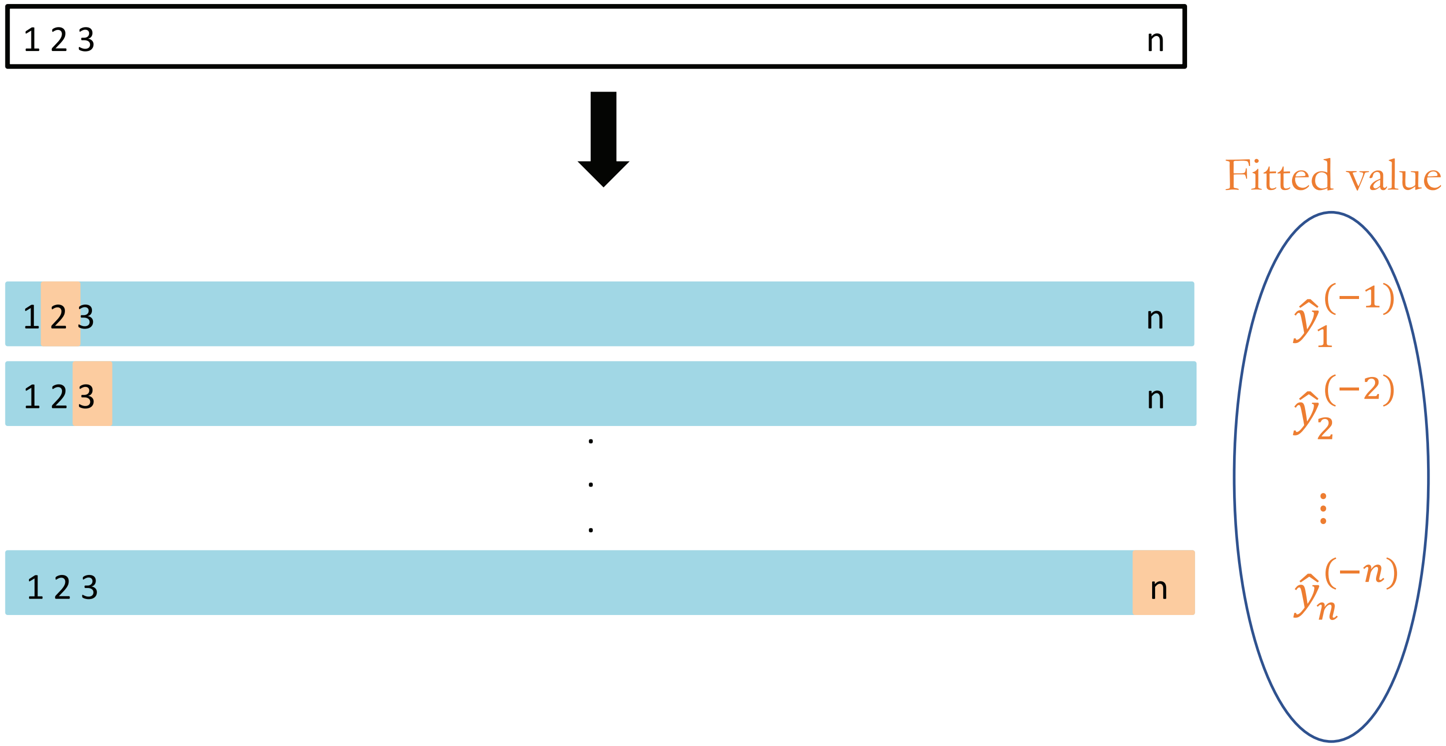
Leave one out cross-validation



Leave one out cross-validation



Leave one out cross-validation



Estimate cross-validation error

Announcements

- Office hours now also available on Mondays and Wednesdays at WVH 208!