Data Mining Techniques

CS 6220 - Section 3 - Fall 2016

Lecture 4: Classification

Jan-Willem van de Meent
(*credit*: Zhao, CS 229, Bishop, Hastie)

Homework Problems

Homework 1 is out on blackboard

- Due: **Friday 30 Sep**
- Can use any language (within reason)
- **Discussion is encouraged, but submissions must be completed individually**
  (absolutely **no** sharing of code)
- Submit as **zip** file ~~via e-mail~~ **on blackboard** by **11.59pm** on day of deadline (no late submissions)
- Please follow *submission guidelines* on website
  (TA's have authority to deduct points)

Classification

Regression Examples

Features → **Continuous Value**

$$\mathbf{x} \Longrightarrow y$$

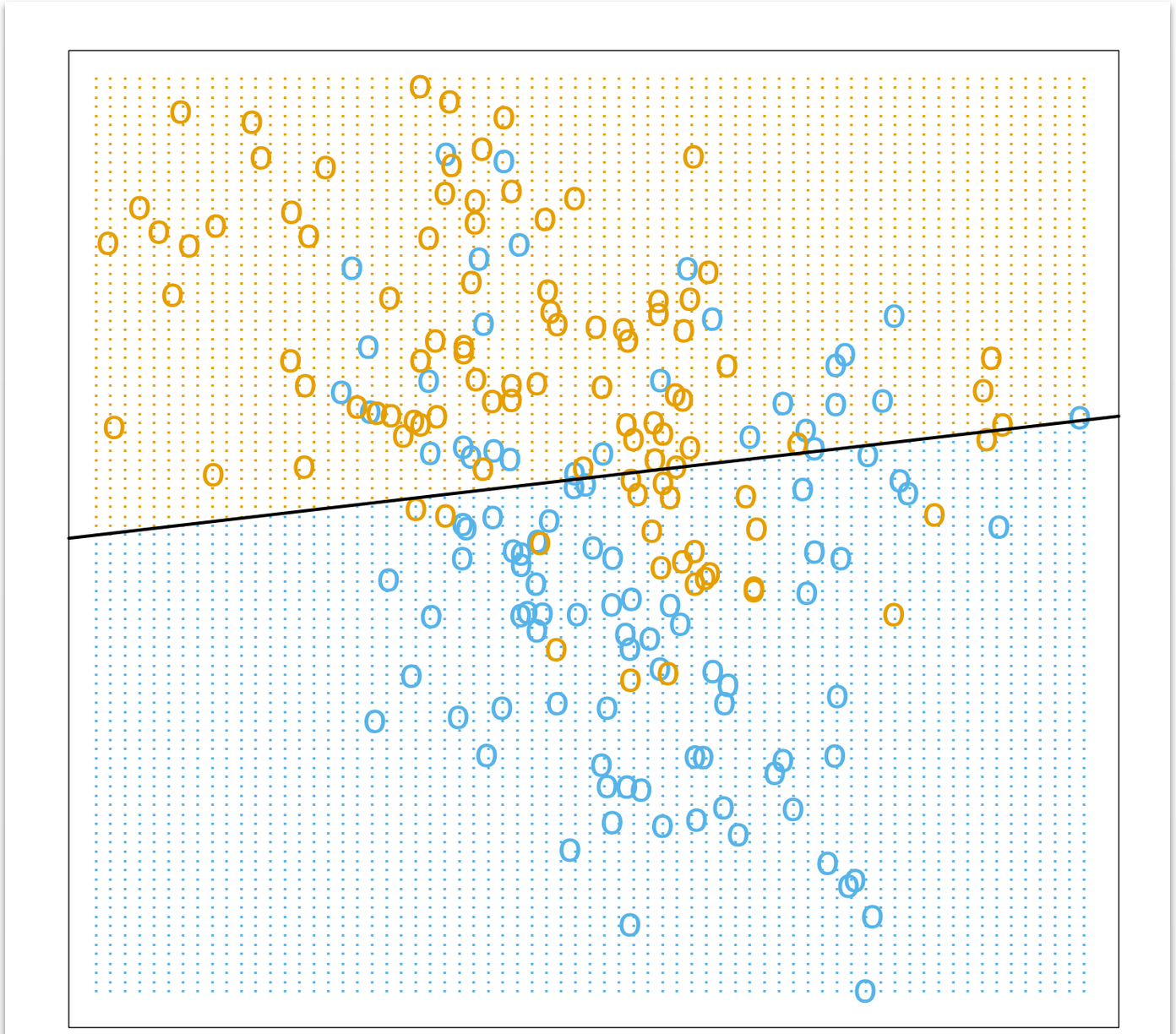
- {age, major, gender, race} ⇒ GPA
- {income, credit score, profession} ⇒ Loan Amount
- {college,major,GPA} ⇒ Future Income

Classification Example

Features		Discrete Value
x	$\Longrightarrow$	y

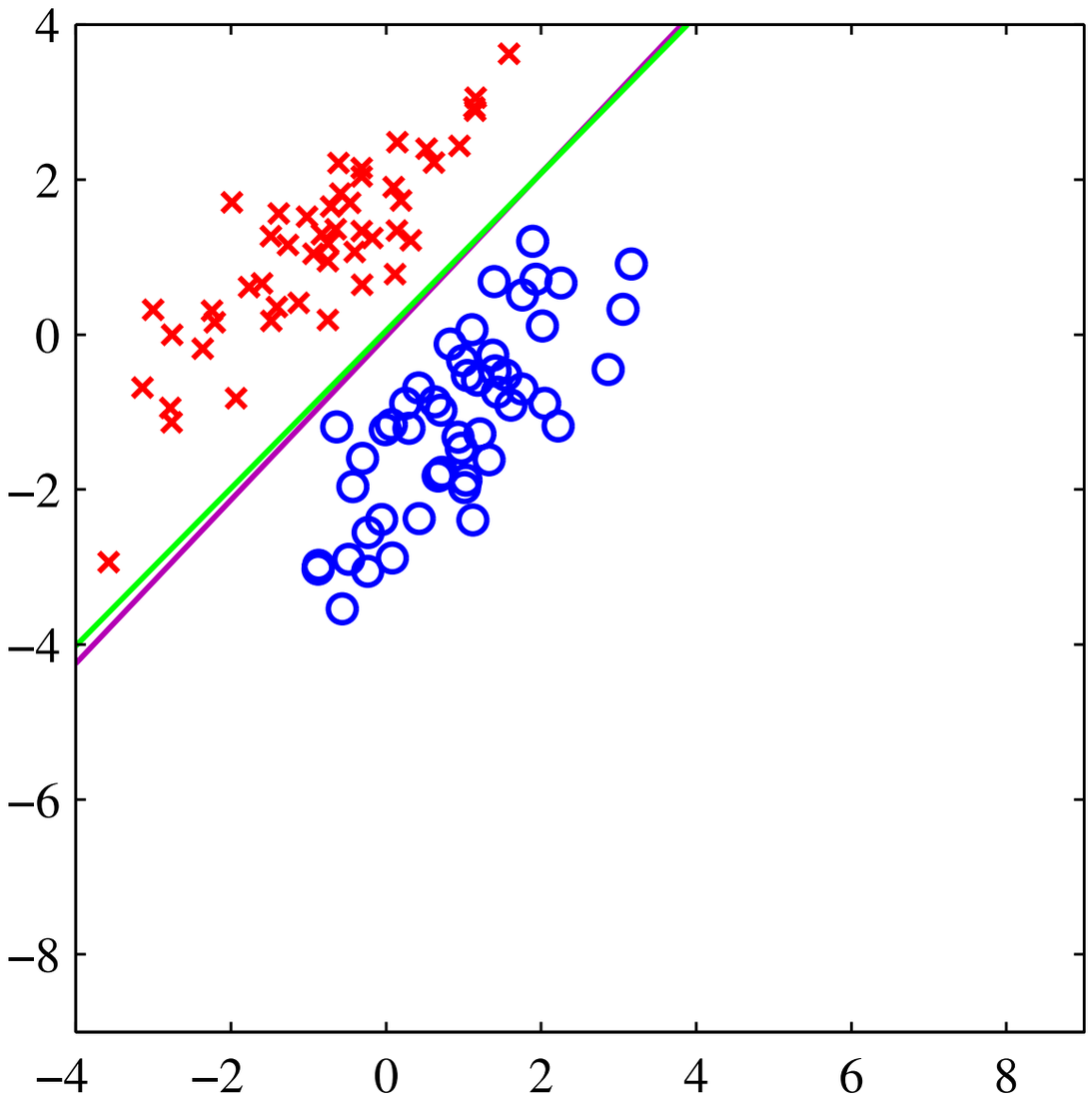
- {words in e-mail} ⇒ SPAM or not?
- {income, credit score, profession} ⇒ Award Loan?
- {pixels in image} ⇒ Cat, dog, or pony?

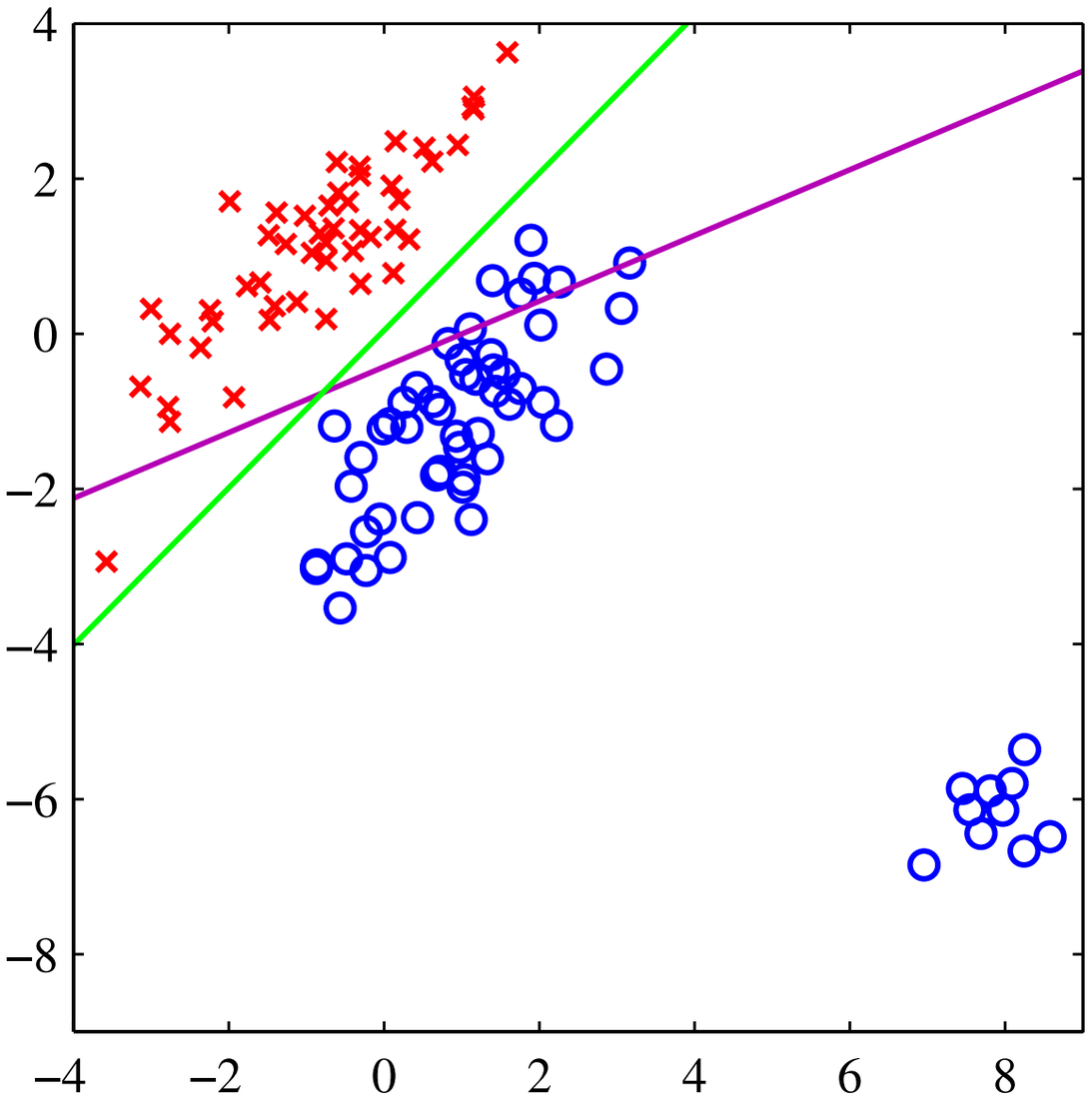
Perceptron (Linear Regression)



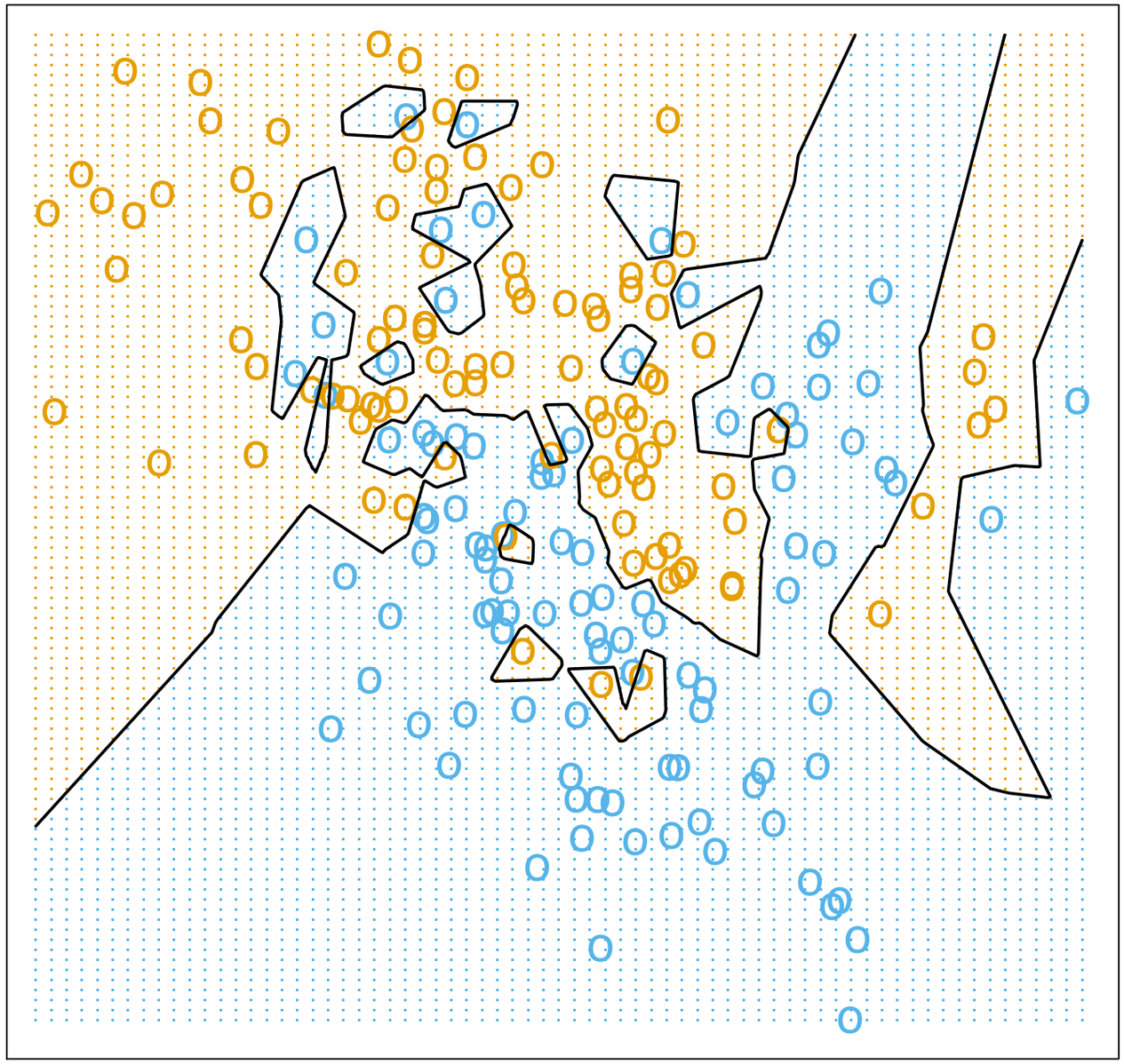
For what types of datasets would this not work well?

Perceptron (Linear Regression)





k-Nearest Neighbors



1 Nearest Neighbor

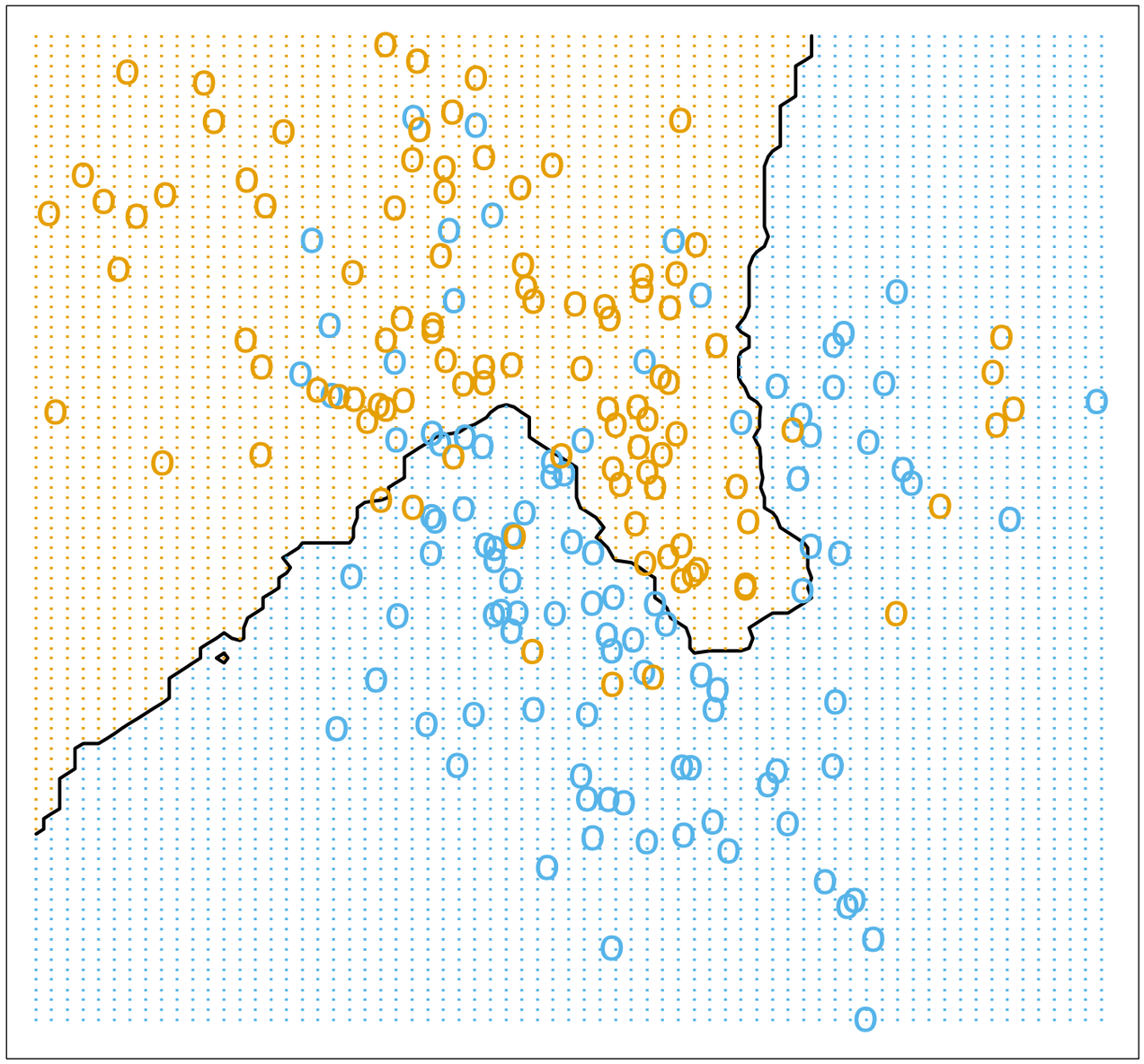
Algorithm

- Find k nearest points to $\boldsymbol{x}^*$
- Predict y^* according to majority vote

Properties

- *Lazy*
  No learned parameters.
- *Instance Learner*
  Needs all data at test time.

k-Nearest Neighbors



15 Nearest Neighbors

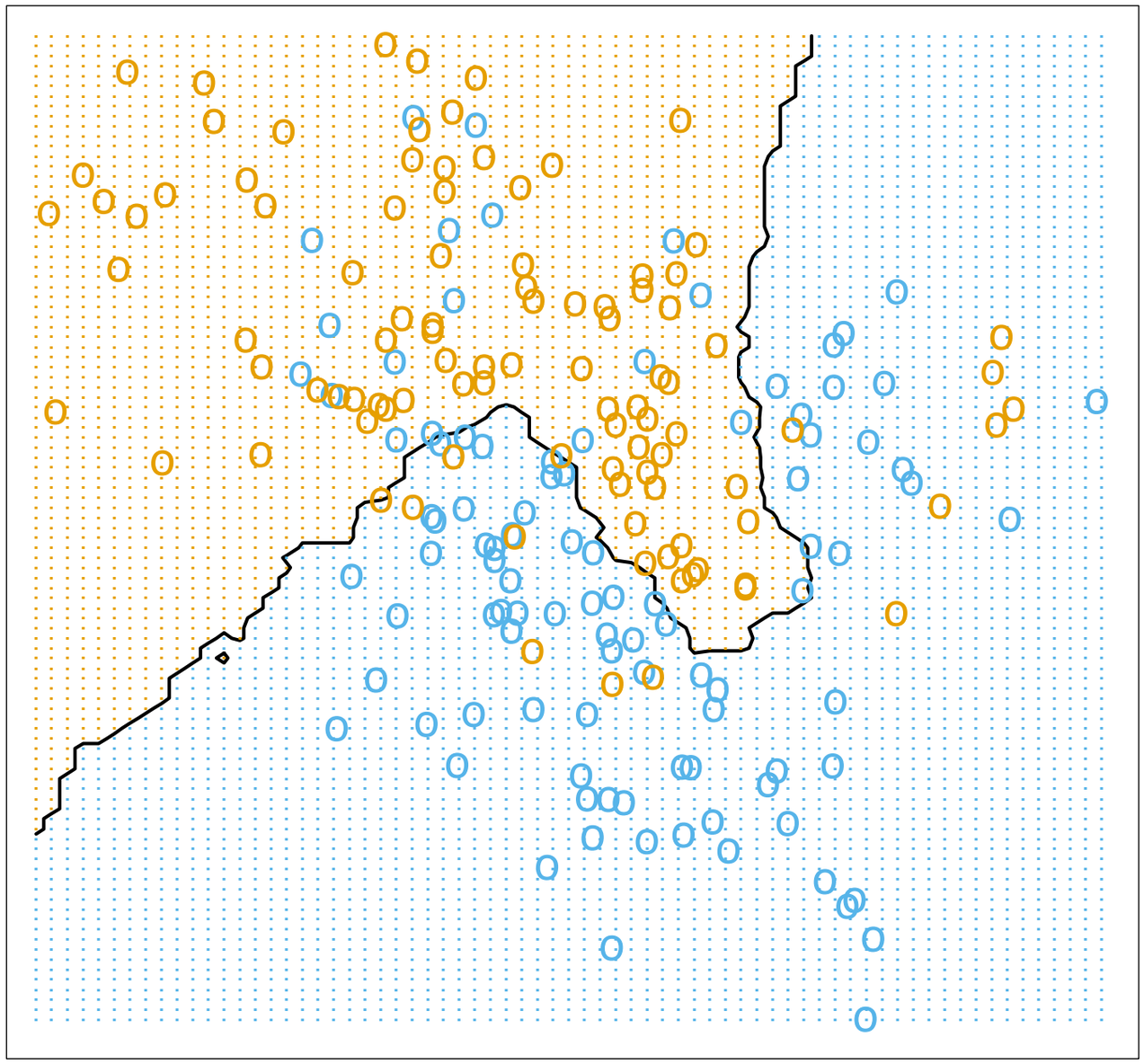
Algorithm

- Find k nearest points to $\boldsymbol{x}^*$
- Predict y^* according to majority vote

Properties

- *Lazy*
  No learned parameters.
- *Instance Learner*
  Needs all data at test time.

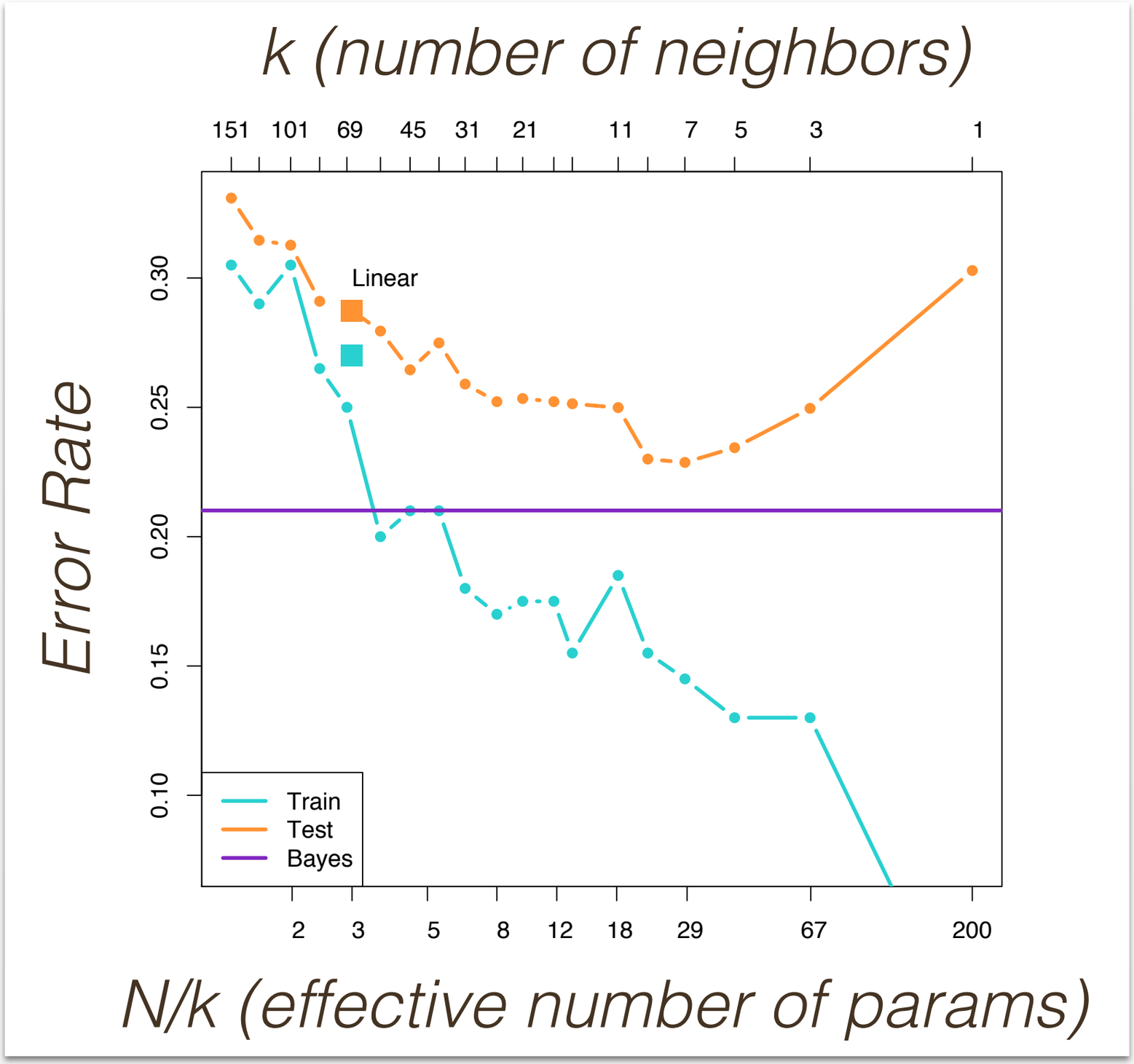
k-Nearest Neighbors



15 Nearest Neighbors

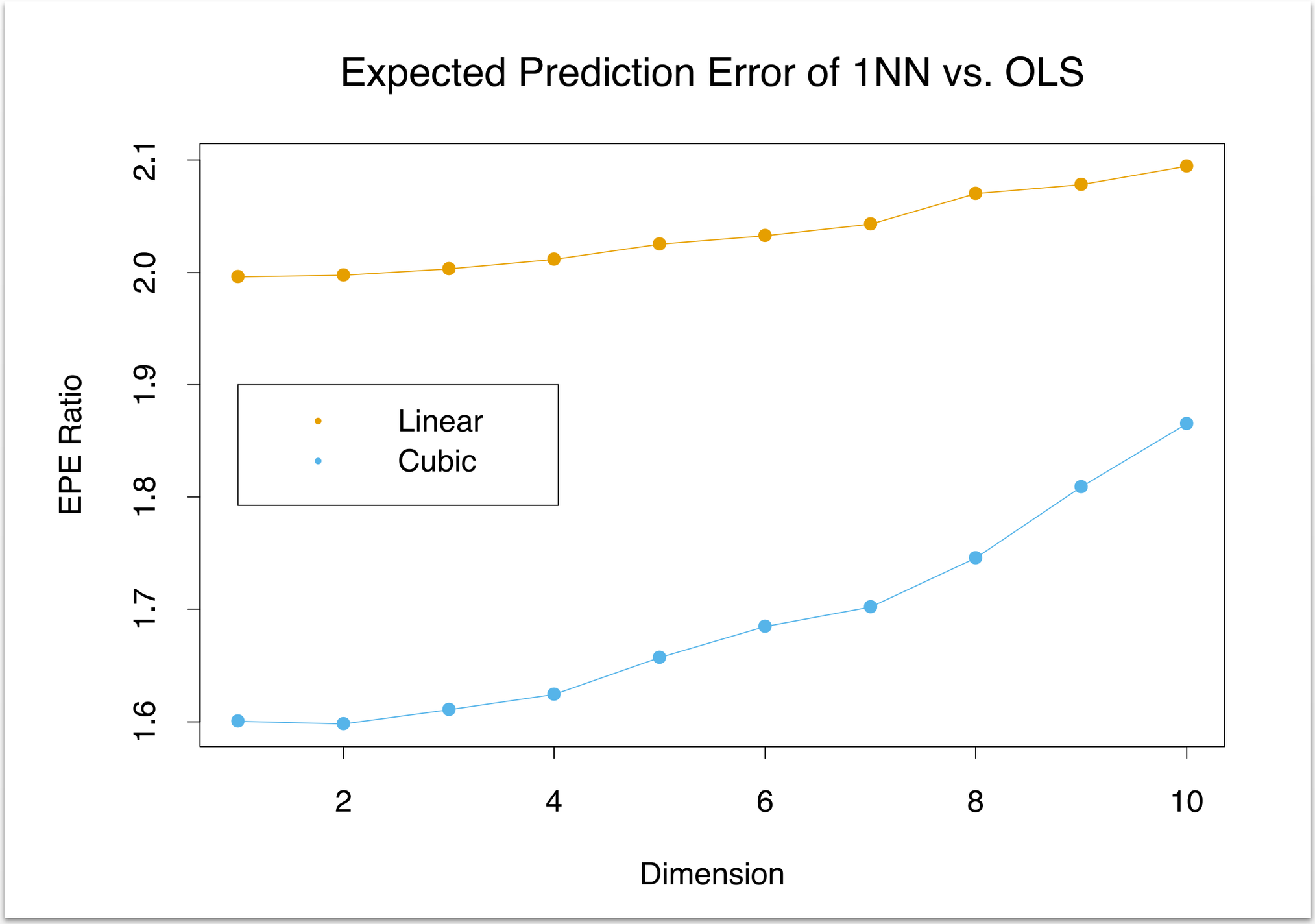
How should we choose k?

Choosing *k*

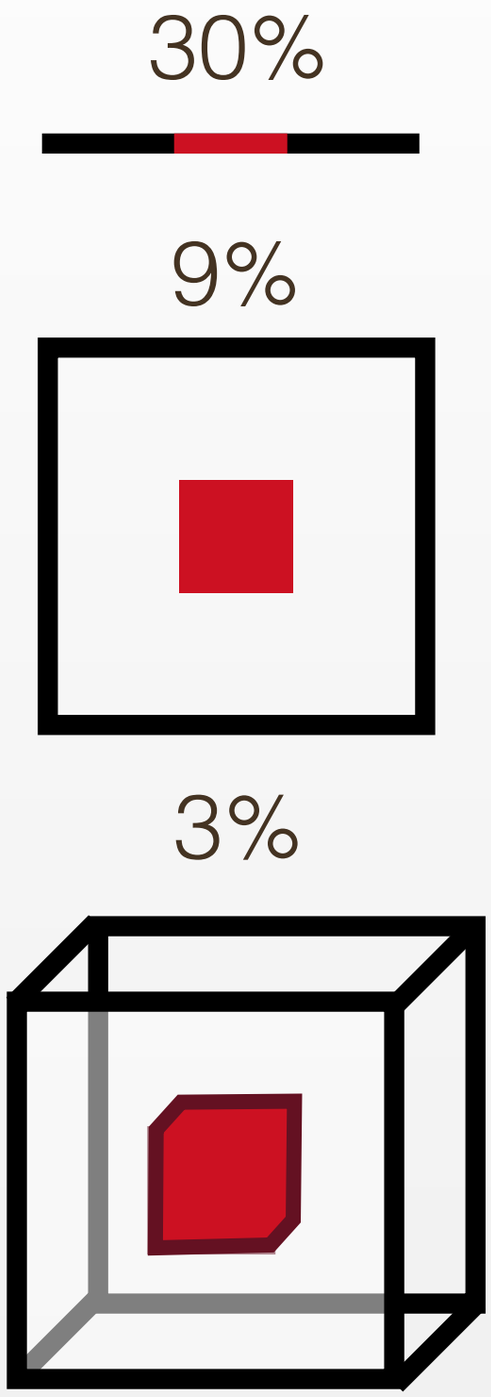

k (number of neighbors)
151
101
69
45
31
21
11
7
5
3
1
Linear
Error Rate
0.30
0.25
0.20
0.15
0.10
Train
Test
Bayes
2
3
5
8
12
18
29
67
200
N/k (effective number of params)

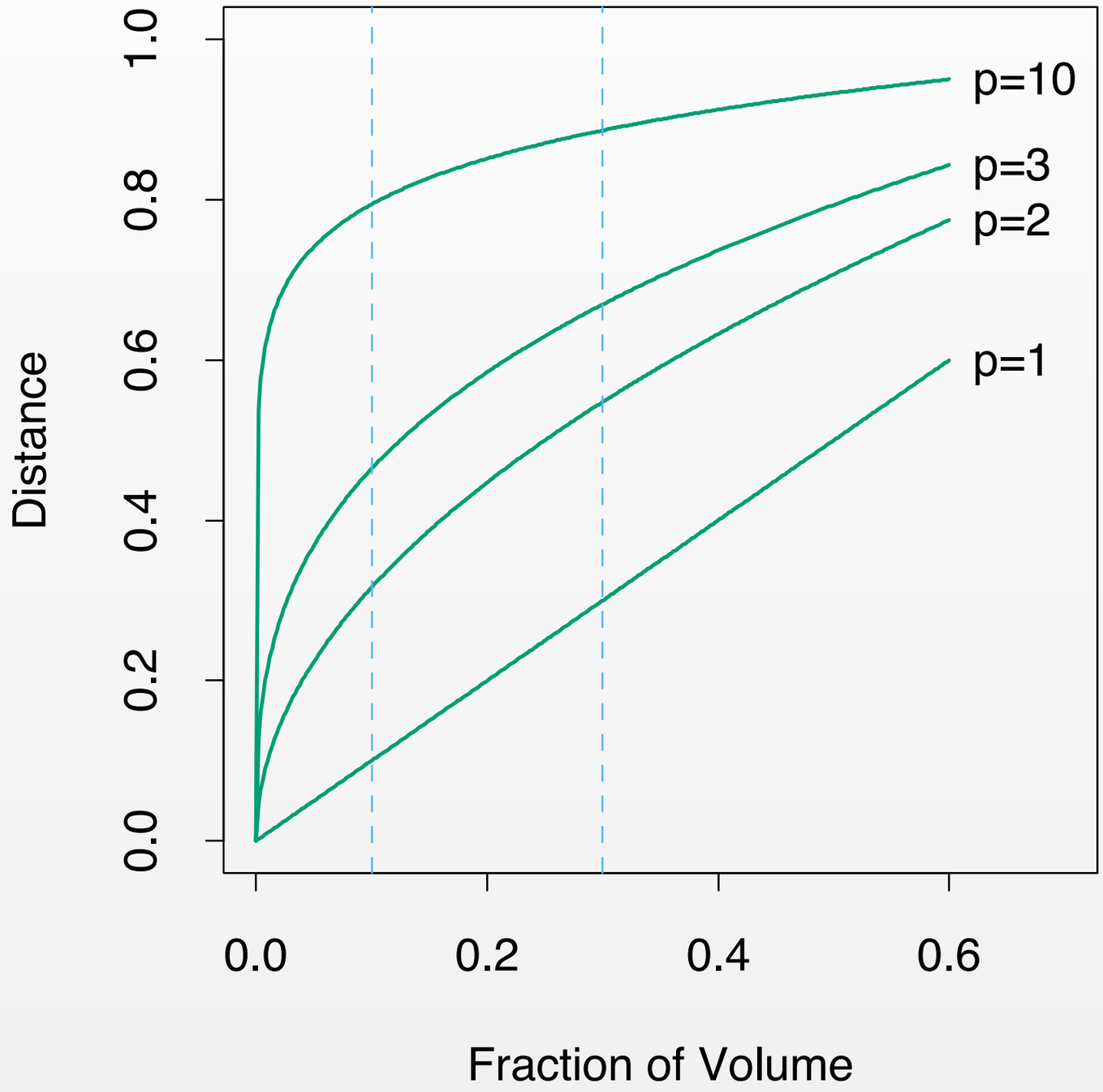
When might k-Nearest Neighbors not perform well?

Curse of Dimensionality

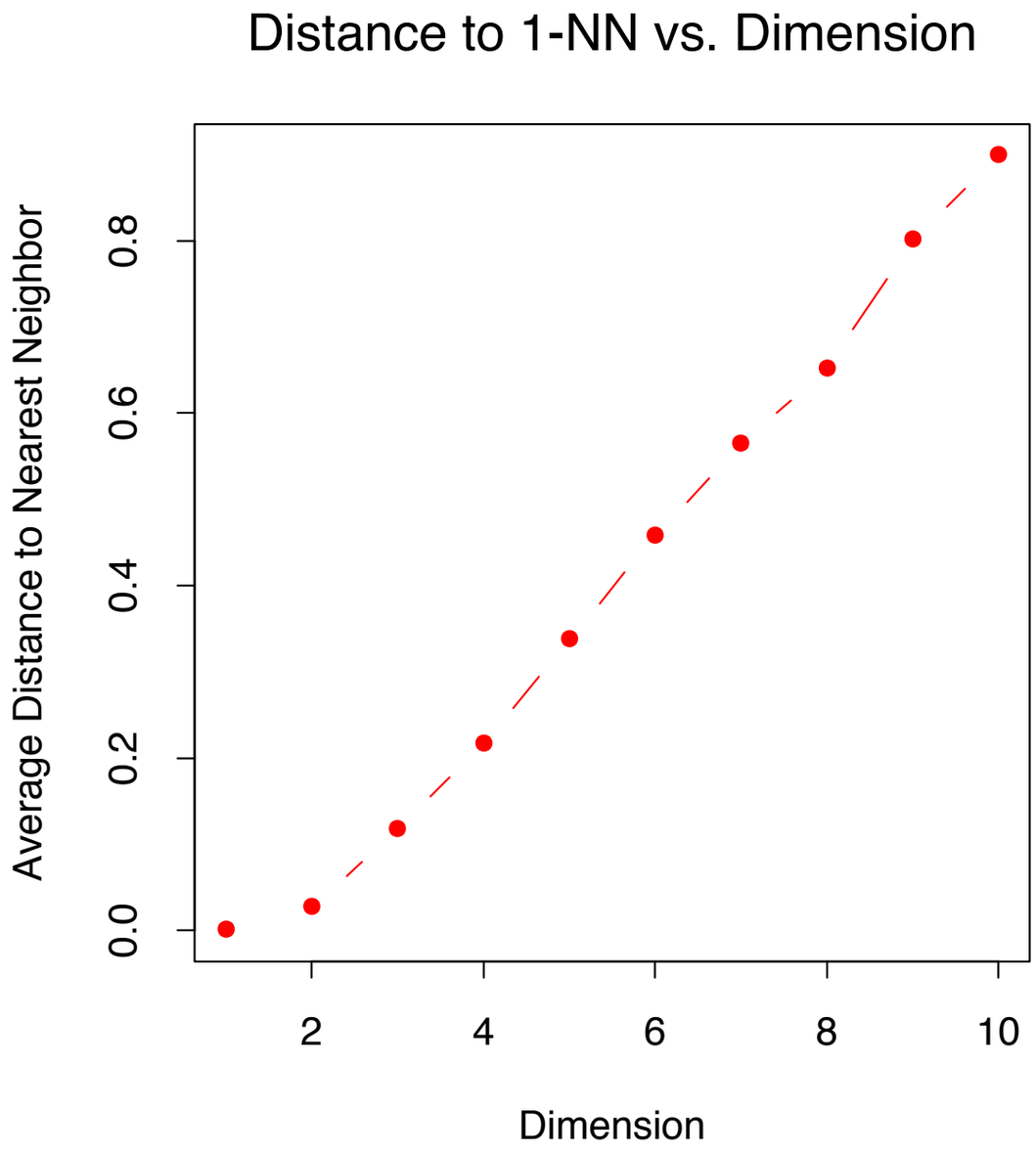


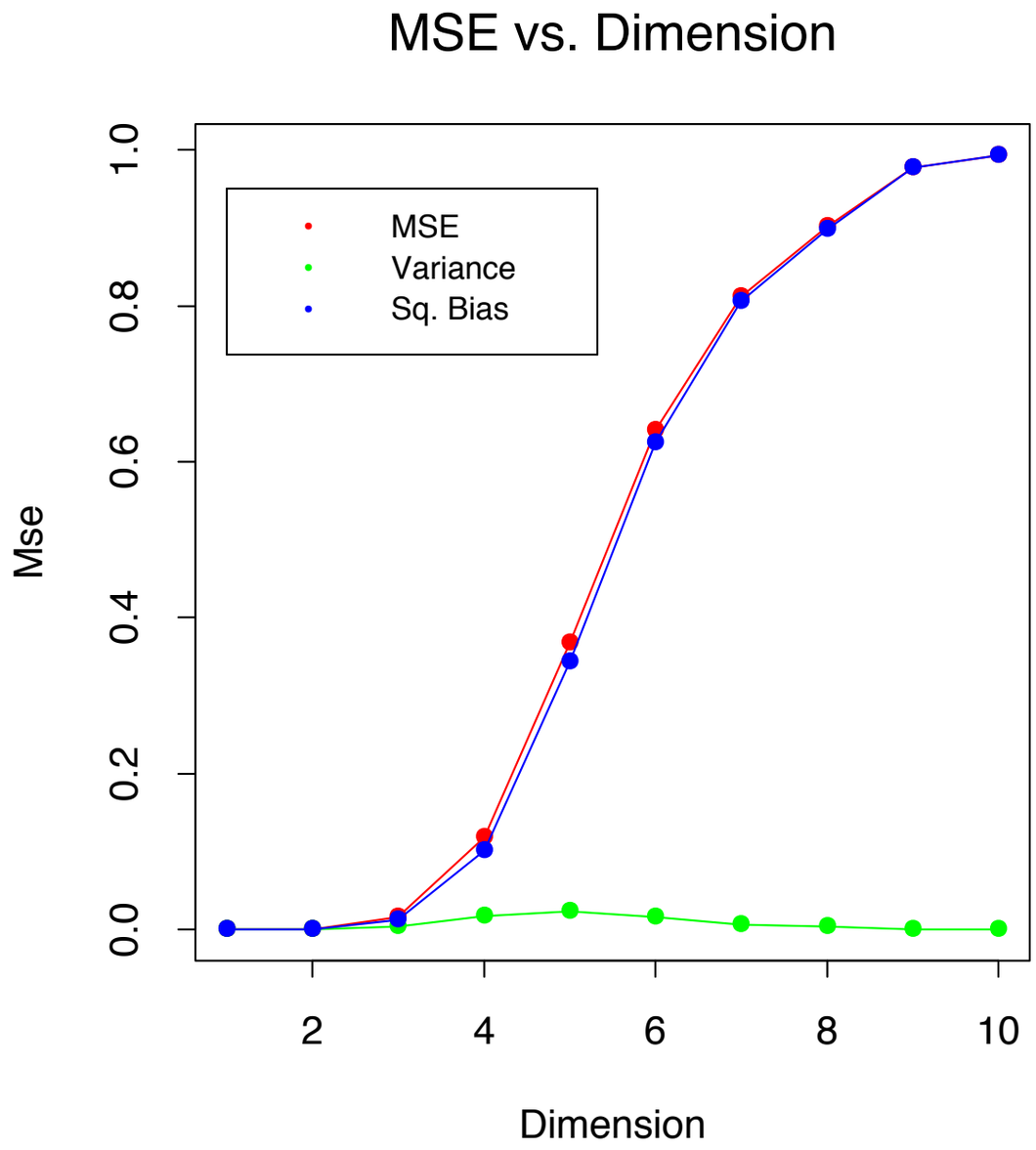
Curse of Dimensionality





Curse of Dimensionality





Distance Norms

- Euclidean Distance $\sqrt{(\sum_{i=1}^{k}(x_i - y_i)^2)}$
- Mahattan Distance $\sum_{i=1}^{k}|x_i - y_i|$
- Minkowski Distance $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}}$

Distance Norms

For nomial variables: Humming Distance

$$D_H(x, y) = \sum_{i=1}^{k} I(x_i, y_i)$$

where

$$I(a, b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{if } a = b \end{cases}$$

Examples:

- $D_H(\text{“}Karolin\text{”}, \text{“}Kathrin\text{”}) = 3$
- $D_H(\text{“}Karolin\text{”}, \text{“}Kerstin\text{”}) = 3$
- $D_H(\text{“}1011101\text{”}, \text{“}1001001\text{”}) = 2$

Sensitivity to Scaling

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Sensitivity to Scaling

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

Normalization Strategies

- Min-Max: $\frac{X - X_{min}}{X_{max} - X_{min}}$

$$a + \frac{(X - X_{min})(b - a)}{X_{max} - X_{min}}$$

- Z-score: $X \sim N(\mu, \sigma)$: $\frac{X - \mu}{\sigma}$

- Scaling: $\frac{X}{X_{max}}$

Linear Methods

Linear Classifiers

$$y_n = f(\boldsymbol{w}^\top \boldsymbol{x}_n)$$

Examples

- Logistic Regression
- Linear Discriminant Analysis
- Naive Bayes

Logistic Regression

$$\log \frac{p(y=k\,|\,\boldsymbol{x})}{p(y=K\,|\,\boldsymbol{x})} = \boldsymbol{w}_k^\top \boldsymbol{x} \qquad k = 1, \ldots, K-1$$

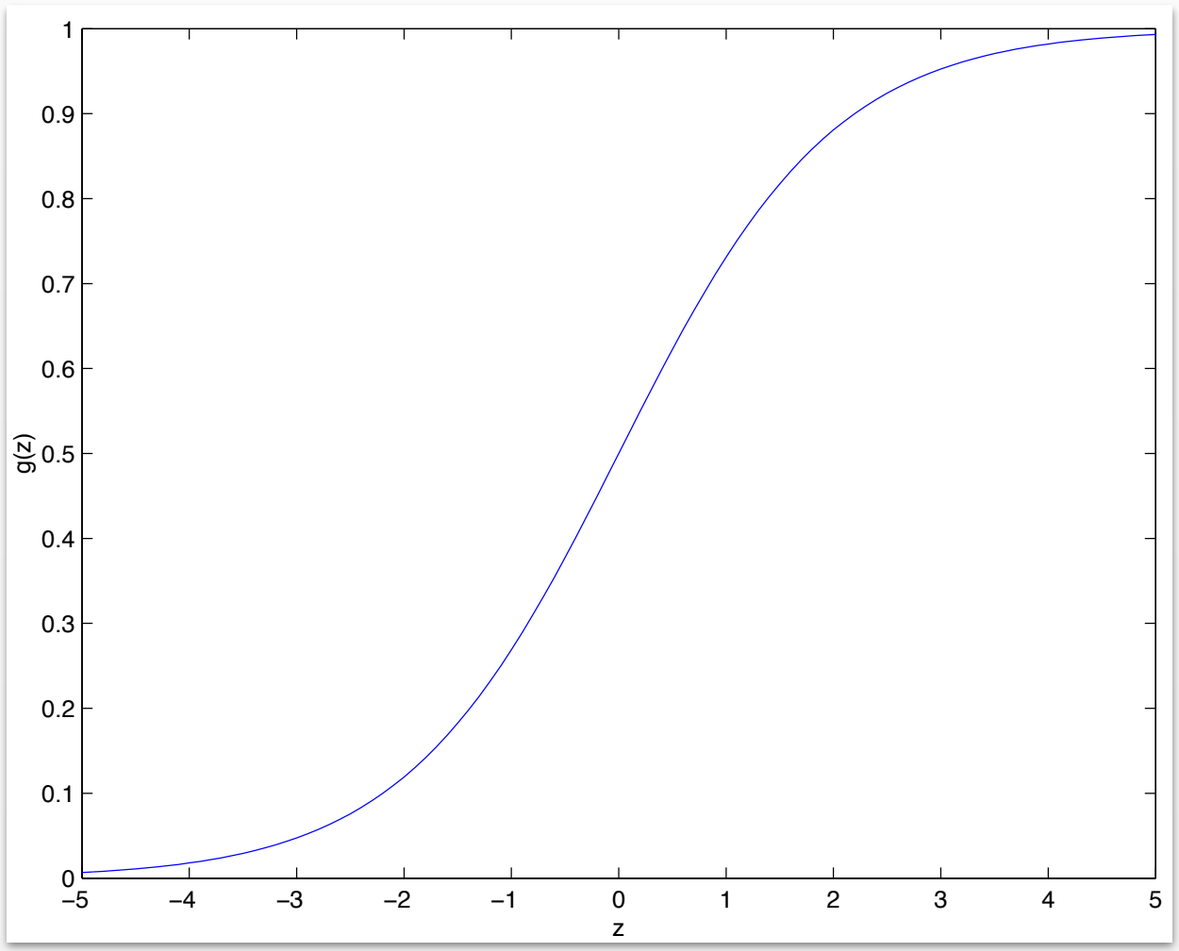
Logistic Regression

$$\log \frac{p(y=k\,|\,\boldsymbol{x})}{p(y=K\,|\,\boldsymbol{x})} = \boldsymbol{w}_k^\top \boldsymbol{x} \qquad k = 1, \ldots, K-1$$

$$\frac{p(y=k\,|\,\boldsymbol{x})}{\sum_l p(y=l\,|\,\boldsymbol{x})} = \frac{\exp^{\boldsymbol{w}_k^\top \boldsymbol{x}}}{1+\sum_{l=1}^{K-1} \exp^{\boldsymbol{w}_l^\top \boldsymbol{x}}} \qquad k = 1, \ldots, K-1$$

$$\frac{p(y=K\,|\,\boldsymbol{x})}{\sum_l p(y=l\,|\,\boldsymbol{x})} = \frac{1}{1+\sum_{l=1}^{K-1} \exp^{\boldsymbol{w}_l^\top \boldsymbol{x}}}$$

Logistic Regression



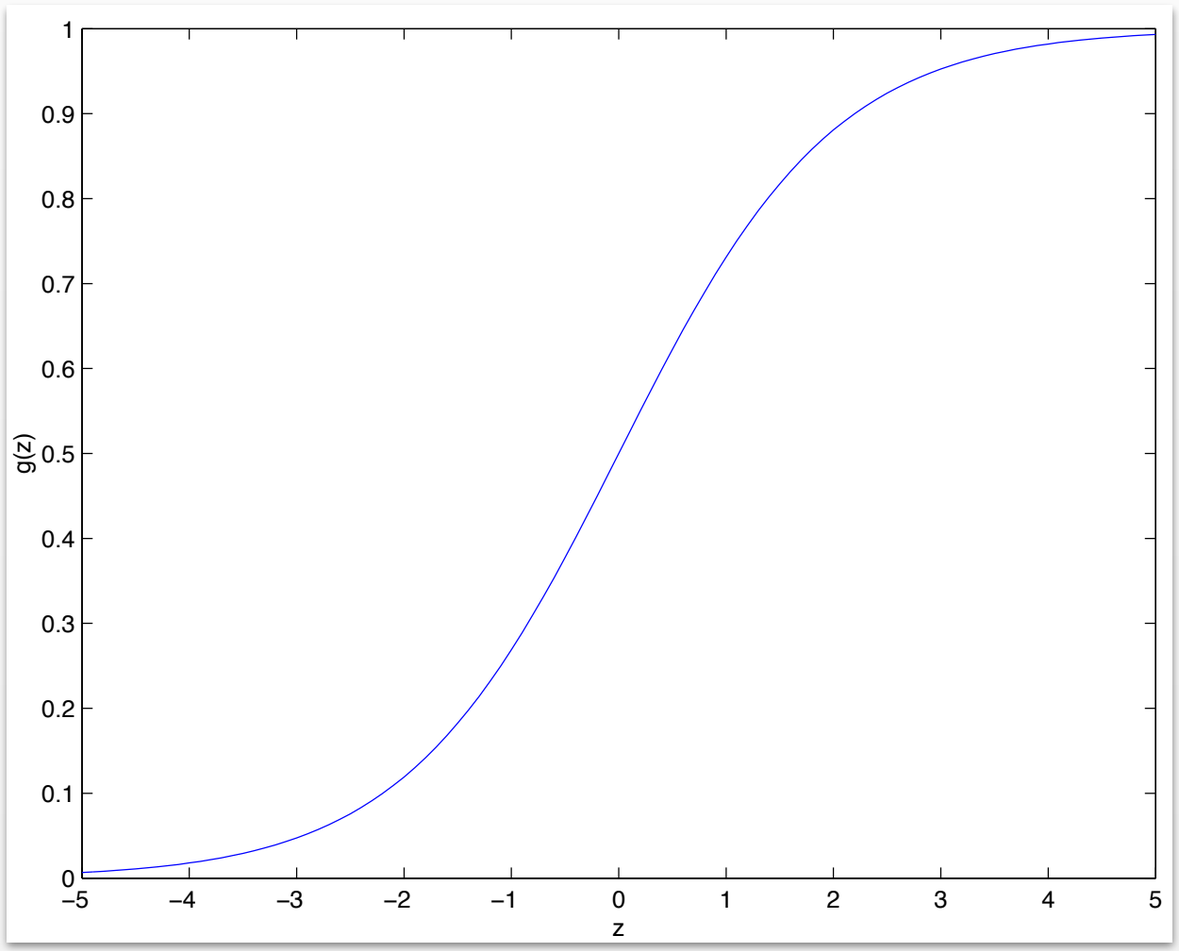
Binary Classification:

$$p(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp^{-\boldsymbol{w}^\top \boldsymbol{x}}}$$

Logistic Function:

$$\sigma(z) = \frac{1}{1 + \exp^{-z}}$$

Logistic Regression



Binary Classification:

$$p(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \exp^{-\boldsymbol{w}^\top \boldsymbol{x}}}$$

Logistic Function:

$$\sigma(z) = \frac{1}{1 + \exp^{-z}}$$

$$y = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$$

Logistic Regression

Maximum Likelihood

$$p(\boldsymbol{y} \mid \boldsymbol{w}) = \prod_{n=1}^{N} \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)^{y_n} (1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n))^{1-y_n}$$

$$\begin{aligned} E(\boldsymbol{w}) &= -\log p(\boldsymbol{y} \mid \boldsymbol{w}) \\ &= -\sum_{n=1}^{N} y_n \log \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) \\ &\quad + (1 - y_n) \log(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)) \end{aligned}$$

Logistic Regression

Gradient

$$\nabla_w E(\boldsymbol{w}) = -\sum_{n=1}^{N} \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_n)\boldsymbol{x}_n \left(\frac{y_n}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} - \frac{1-y_n}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} \right)$$

$$\sigma(z) = \frac{1}{1+\exp^{-z}} \qquad \sigma'(z) = ?$$

Logistic Regression

Gradient

$$\nabla_w E(\boldsymbol{w}) = -\sum_{n=1}^{N} \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_n)\boldsymbol{x}_n \left(\frac{y_n}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} - \frac{1-y_n}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} \right)$$

$$\sigma(z) = \frac{1}{1+\exp^{-z}}$$

$$\begin{aligned}
\sigma'(z) &= \frac{1}{(1+\exp^{-z})^2} \exp^{-z} \\
&= \frac{1}{(1+\exp^{-z})} \frac{\exp^{-z}}{(1+\exp^{-z})} \\
&= \sigma(z)(1-\sigma(z))
\end{aligned}$$

Logistic Regression

Gradient

$$\nabla_w E(\boldsymbol{w}) = -\sum_{n=1}^{N} \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_n)\boldsymbol{x}_n \left(\frac{y_n}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} - \frac{1-y_n}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} \right)$$

$$\sigma(z) = \frac{1}{1+\exp^{-z}} \qquad \sigma'(z) = \sigma(z)(1-\sigma(z))$$

Logistic Regression

Gradient

$$\nabla_w E(\boldsymbol{w}) = -\sum_{n=1}^{N} \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_n)\boldsymbol{x}_n \left(\frac{y_n}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} - \frac{1-y_n}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} \right)$$

Logistic Regression

Gradient

$$\nabla_w E(\boldsymbol{w}) = -\sum_{n=1}^{N} \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_n)\boldsymbol{x}_n \left(\frac{y_n}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} - \frac{1-y_n}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} \right)$$

$$= -\sum_{n=1}^{N} \boldsymbol{x}_n \left((1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n))y_n - (1-y_n)\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) \right)$$

Logistic Regression

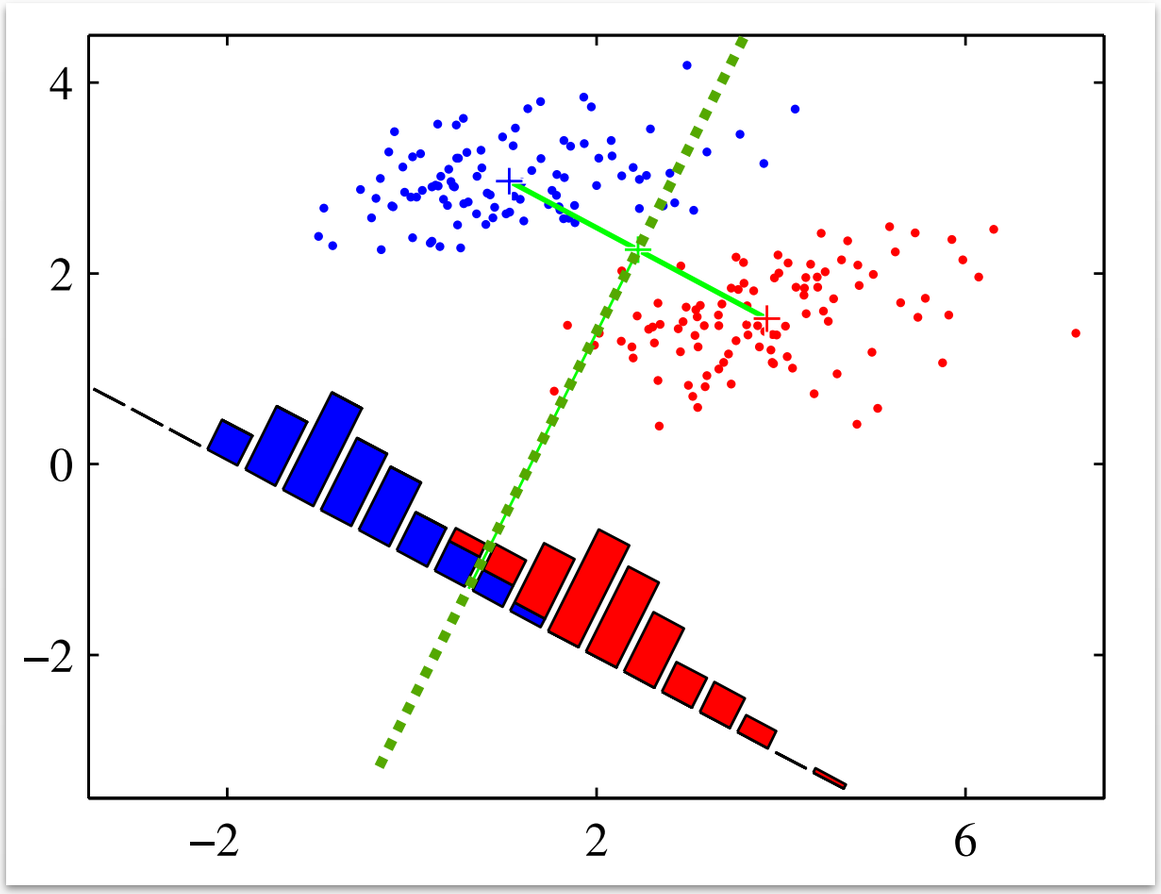
Gradient

$$\nabla_w E(\boldsymbol{w}) = -\sum_{n=1}^{N} \sigma'(\boldsymbol{w}^\top \boldsymbol{x}_n)\boldsymbol{x}_n \left(\frac{y_n}{\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} - \frac{1-y_n}{1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)} \right)$$

$$= -\sum_{n=1}^{N} \boldsymbol{x}_n \left((1-\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n))y_n - (1-y_n)\sigma(\boldsymbol{w}^\top \boldsymbol{x}_n) \right)$$

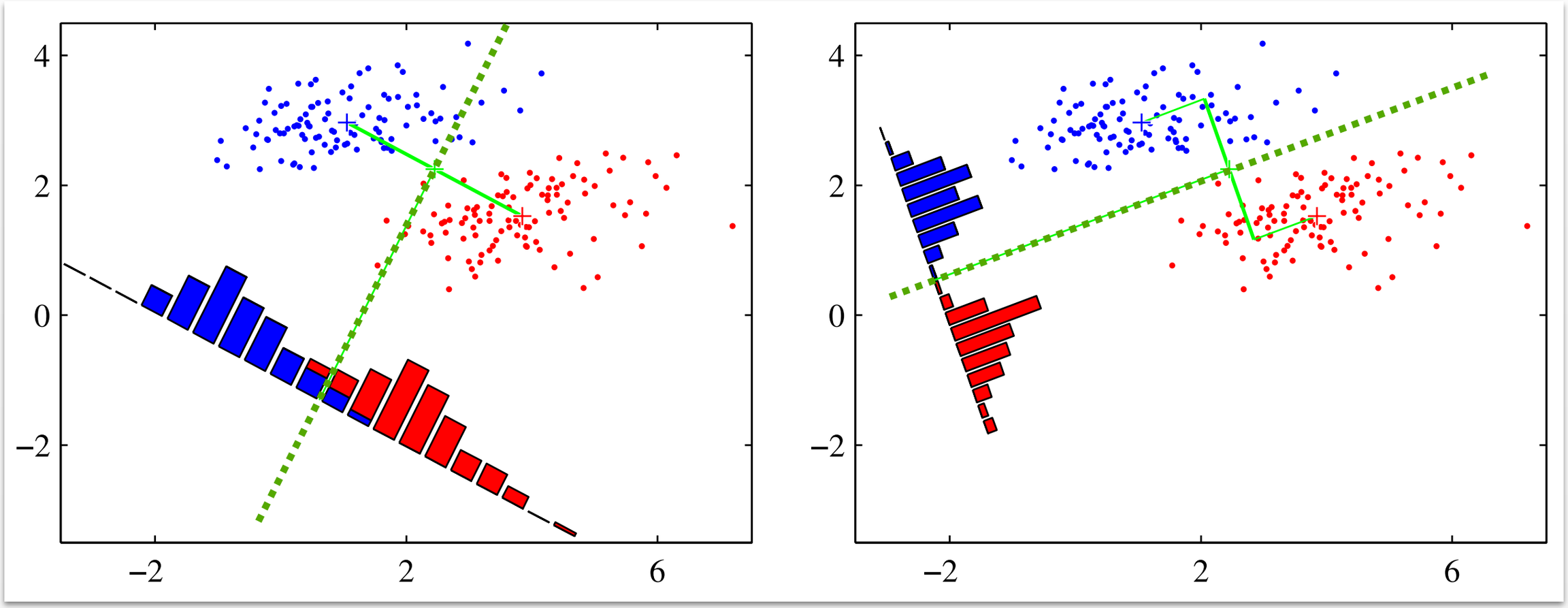
$$= -\sum_{n=1}^{N} \boldsymbol{x}_n (y_n - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n))$$

Linear Discriminant Analysis

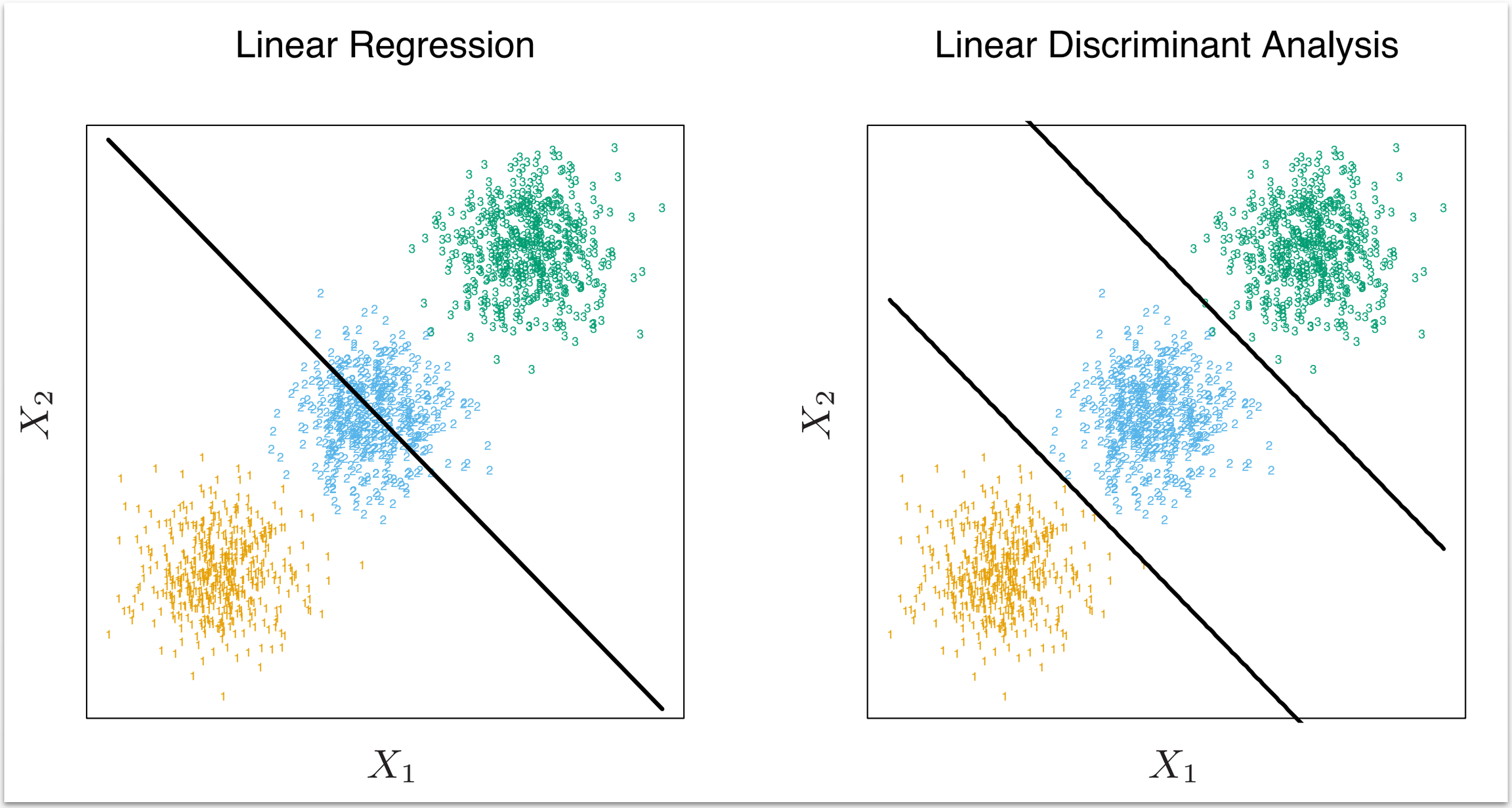


Idea: Calculate center for each class

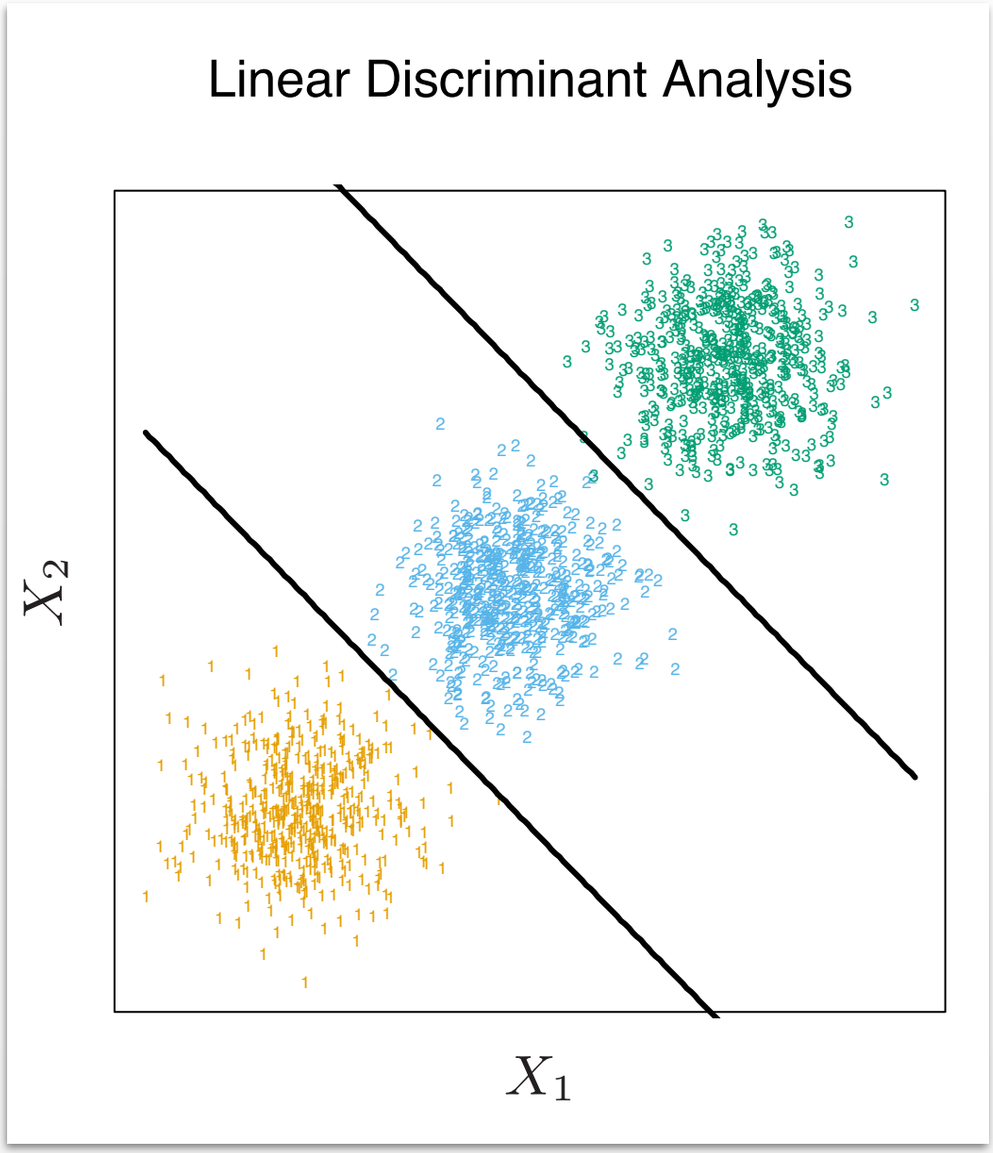
Linear Discriminant Analysis



Linear Discriminant Analysis



Linear Discriminant Analysis



Algorithm

- Mean for each class

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n:y_n=k} \boldsymbol{x}_n$$

- Covariance for each class

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n:y_n=k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

- Average covariance

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_k N_k \Sigma_k$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \mid y = k)$$

Linear Discriminant Analysis

$$p(\boldsymbol{x} \mid y = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

Quadratic Discriminant Analysis

$$p(\boldsymbol{x} \mid y = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Algorithm

- Mean for each class

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n:y_n=k} \boldsymbol{x}_n$$

- Covariance for each class

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n:y_n=k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top (\boldsymbol{x}_n - \boldsymbol{\mu}_k)$$

- Average covariance

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_k N_k \Sigma_k$$

Linear Discriminant Analysis

Predict using likelihood

$$y^* = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}^* \mid y = k)$$

Linear Discriminant Analysis

$$p(\boldsymbol{x} \mid y = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

Quadratic Discriminant Analysis

$$p(\boldsymbol{x} \mid y = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

What is “linear” or “quadratic” about these methods?

Linear Discriminant Analysis

$$\log \frac{p(\boldsymbol{x} \mid y = k)}{p(\boldsymbol{x} \mid y = l)} = \frac{1}{2}(\boldsymbol{x} - \mu_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu_l)$$
$$- \frac{1}{2}(\boldsymbol{x} - \mu_k)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu_k)$$

Linear Discriminant Analysis

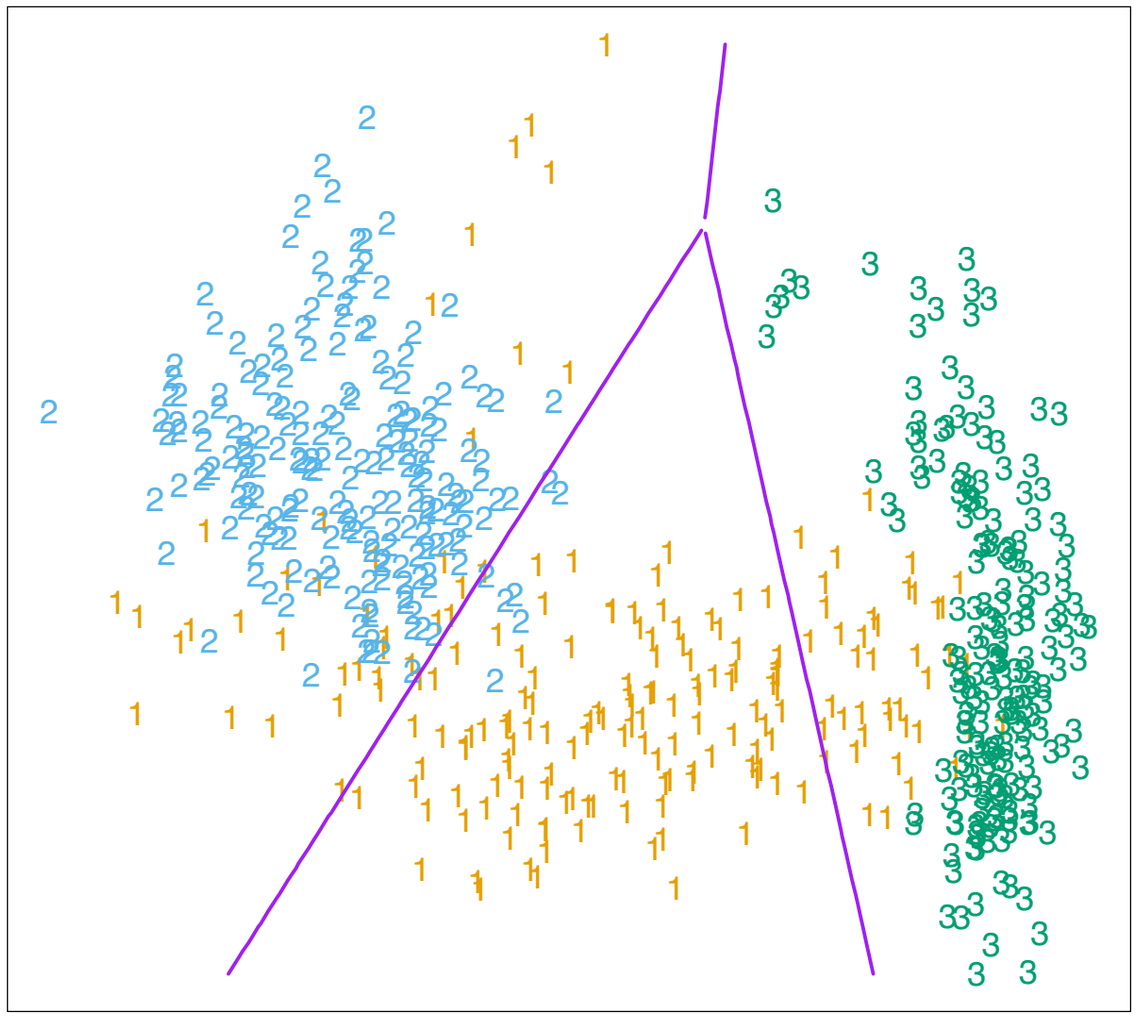
$$
\begin{aligned}
\log \frac{p(\boldsymbol{x} \mid y=k)}{p(\boldsymbol{x} \mid y=l)} &= \frac{1}{2}(\boldsymbol{x}-\mu_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu_l) \\
&\quad - \frac{1}{2}(\boldsymbol{x}-\mu_k)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu_k) \\
&= -\frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_l)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l) \\
&\quad + \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)
\end{aligned}
$$

Logistic Regression vs LDA

- Both methods use linear log odds
- *LDA*: Calculate weights directly using mean and covariance for each class
  - Easy to calculate
  - Can be more sensitive to outliers
- *Logistic Regression*: Perform maximum likelihood estimation with logistic activation function
  - Harder to solve, but more expressive.

Linear Discriminant Analysis

LDA



QDA