ECE595 / STAT598: Machine Learning I
Lecture 13 Connecting Bayesian with Linear Regression

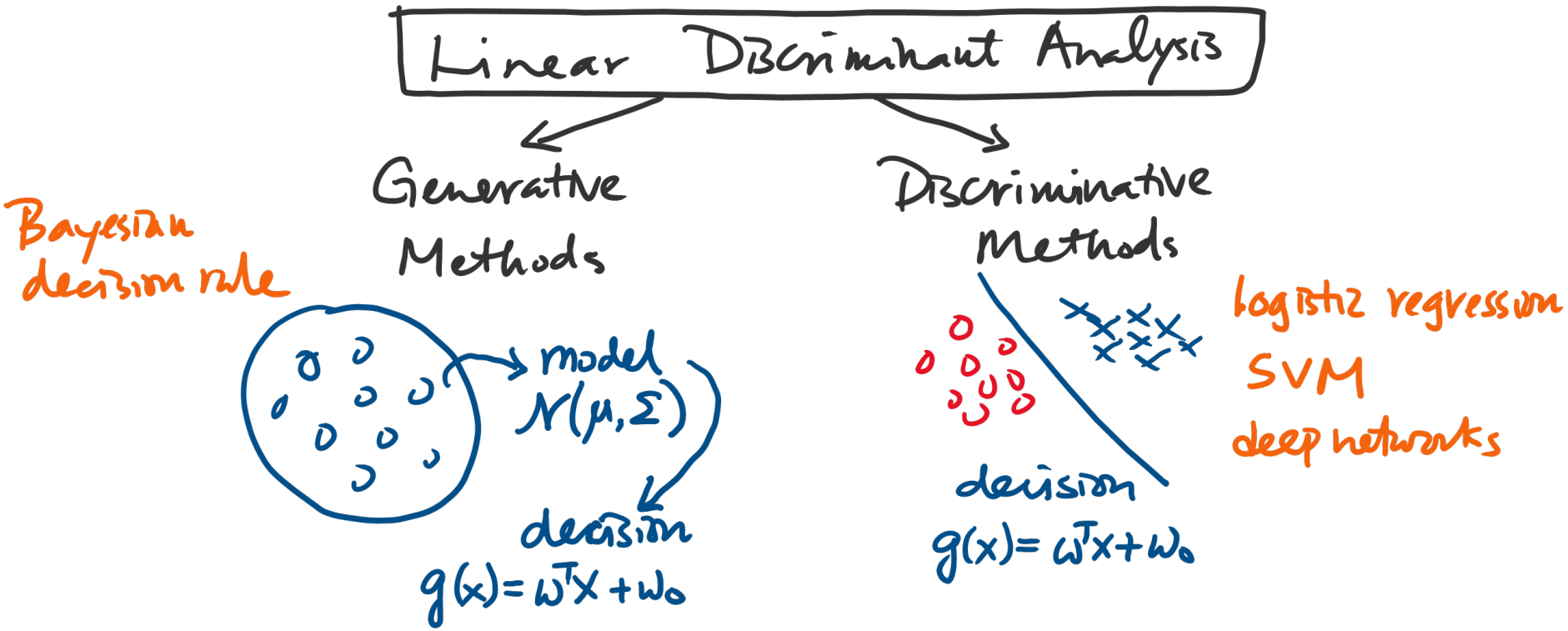
Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Overview



- In linear discriminant analysis (LDA), there are generally two types of approaches
- **Generative approach**: Estimate model, then define the classifier
- **Discriminative approach**: Directly define the classifier

Outline

Generative Approaches

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Linear Regression Review
  - Linear regression in the context of classification
  - Linking linear regression with MLE and MAP
- Connection between Linear Regression and Bayesian
  - Expected Loss
  - Main Result
  - Implications

Linear Regression Reviewed

- Linear regression is actually a **discriminative method**.
- Do not require a distributional model.
- Construct the hypothesis function directly:

$$h(\boldsymbol{x}) = \begin{cases} +1, & \text{if} \quad g(\boldsymbol{x}) > 0, \\ -1, & \text{if} \quad g(\boldsymbol{x}) < 0. \end{cases}$$

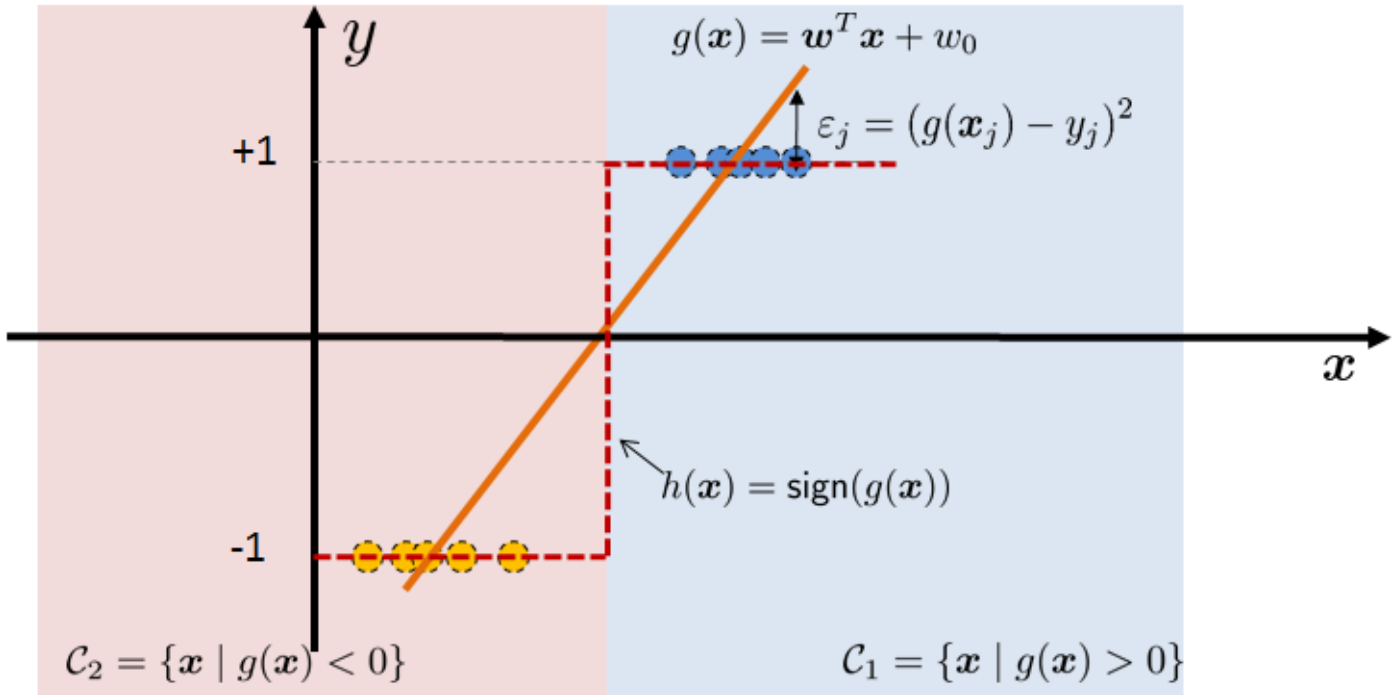
- Consider a binary classification problem with discriminant function:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

- The goal is to determine the parameters $\boldsymbol{\theta} = \{\boldsymbol{w}, w_0\}$
- Training data: $(\boldsymbol{x}_n, y_n)_{n=1}^N$
  - $\boldsymbol{x}_n \in \mathbb{R}^d$ is the input vector
  - $y_n \in \{-1, +1\}$ is the corresponding label

Geometry of Linear Regression

- The discriminant function $g(\boldsymbol{x})$ is linear
- The hypothesis function $h(\boldsymbol{x}) = \text{sign}(g(\boldsymbol{x}))$ is a unit step



Loss Function

- All discriminant algorithms have a **Training Loss Function**

$$J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(g(\boldsymbol{x}_n), y_n).$$

- In linear regression,

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{n=1}^{N}(g(\boldsymbol{x}_n) - y_n)^2 \\ &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_n + w_0 - y_n)^2 \\ &= \frac{1}{N}\left\|\begin{bmatrix}\boldsymbol{x}_1^T & 1\\ \vdots & \vdots \\ \boldsymbol{x}_N^T & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\ w_0\end{bmatrix} - \begin{bmatrix}y_1\\ \vdots \\ y_N\end{bmatrix}\right\|^2 = \frac{1}{N}\|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2. \end{aligned}$$

Solution of Linear Regression

Theorem (Linear Regression Solution)

The loss function of a linear regression model is given by

$$J(\boldsymbol{\theta}) = \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2,$$

of which the minimizer is

$$\boldsymbol{\theta}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}.$$

- Take derivative and setting to zero:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \right\} \\ &= 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}) = \boldsymbol{0}.\end{aligned}$$

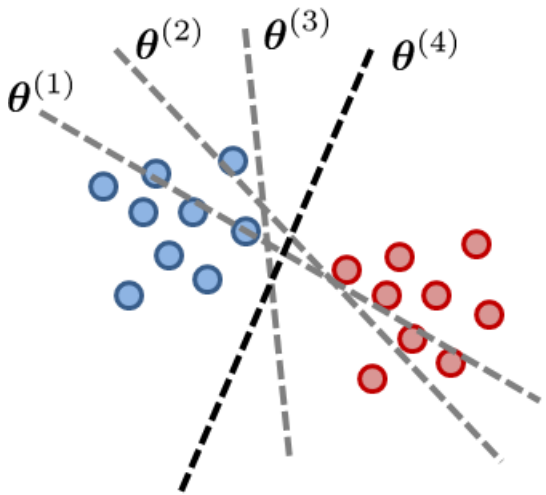
- So solution is $\boldsymbol{\theta}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$, assuming $\boldsymbol{A}^T\boldsymbol{A}$ is invertible.

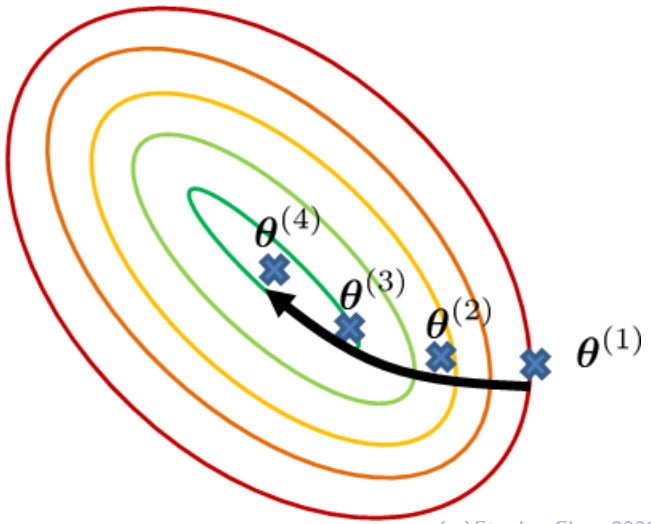
When $\boldsymbol{A}^T\boldsymbol{A}$ is large

- Computing $(\boldsymbol{A}^T\boldsymbol{A})^{-1}$ directly is infeasible for large-scale datasets with a large number of variables
- Consider using iterative algorithms such as gradient descent
- The gradient descent is given by the iteration:

$$\begin{aligned}\boldsymbol{\theta}^{(k+1)} &= \boldsymbol{\theta}^{(k)} - \eta\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^{(k)}) \\ &= \boldsymbol{\theta}^{(k)} - \eta(2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\theta}^{(k)} - 2\boldsymbol{A}^T\boldsymbol{y})\end{aligned}$$

- A pictorial illustration of the gradient descent step:





Treating Linear Regression as Maximum-Likelihood

- Minimizing $J(\boldsymbol{\theta})$ is the same as solving a **maximum-likelihood**:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \sum_{n=1}^{N} (\boldsymbol{a}_n^T \boldsymbol{\theta} - y_n)^2 \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \exp\left\{ -\sum_{n=1}^{N} (\boldsymbol{a}_n^T \boldsymbol{\theta} - y_n)^2 \right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\boldsymbol{a}_n^T \boldsymbol{\theta} - y_n)^2}{2\sigma^2} \right\} \right\}
\end{aligned}$$

- Assume noise is i.i.d. Gaussian with variance σ^2.

Treating Linear Regression as Maximum-a-Posteriori

- We can modify the MLE by adding a prior

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \exp\left\{ -\frac{\rho(\boldsymbol{\theta})}{\beta} \right\}.$$

- Then, we have a MAP problem:

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\text{argmax}} \quad \prod_{n=1}^{N} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(\boldsymbol{a}_n^T\boldsymbol{\theta} - y_n)^2}{2\sigma^2} \right\} \right\} \exp\left\{ -\frac{\rho(\boldsymbol{\theta})}{\beta} \right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\boldsymbol{a}_n^T\boldsymbol{\theta} - y_n)^2 + \frac{1}{\beta}\rho(\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \quad \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\rho(\boldsymbol{\theta}), \qquad \text{where} \quad \lambda = 2\sigma^2/\beta.
\end{aligned}$$

- $\rho(\cdot)$ is called **regularization function**.
- Useful when $\boldsymbol{A}^T\boldsymbol{A}$ is not invertible.

Ridge Regression

- One option: Choose a Gaussian prior

$$\exp\left\{-\frac{\rho(\boldsymbol{\theta})}{\beta}\right\} = \exp\left\{-\frac{\|\boldsymbol{\theta}\|^2}{2\sigma_0^2}\right\}$$

- Then, the MAP becomes

$$\begin{aligned}
\boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{n=1}^{N} \left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\boldsymbol{a}_n^T\boldsymbol{\theta} - y_n)^2}{2\sigma^2}\right\}\right\} \exp\left\{-\frac{\|\boldsymbol{\theta}\|^2}{2\sigma_0^2}\right\} \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{n=1}^{N} (\boldsymbol{a}_n^T\boldsymbol{\theta} - y_n)^2 + \underbrace{\frac{\sigma^2}{\sigma_0^2}}_{=\lambda} \|\boldsymbol{\theta}\|^2 \\
&= \underset{\boldsymbol{\theta}}{\text{argmin}} \|\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2
\end{aligned}$$

- This is called **Tikhonov regularization** or **Ridge regression**.

Outline

Generative Approaches

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Linear Regression Review
  - Linear regression in the context of classification
  - Linking linear regression with MLE and MAP
- Connection between Linear Regression and Bayesian
  - Expected Loss
  - Main Result
  - Implications

Connection with Bayesian Decision Rule

- With infinite training samples, $J(\boldsymbol{\theta})$ converges almost surely to its expectation

$$\frac{1}{N}\sum_{n=1}^{N}(g(\boldsymbol{x}_n)-y_n)^2 \xrightarrow{p} \mathbb{E}_{\boldsymbol{x},y}[g(\boldsymbol{x})-y)^2].$$

- Minimizing $J(\boldsymbol{\theta})$ is essentially minimizing the expectation

$$\begin{aligned}\boldsymbol{\theta}^* &= \underset{\boldsymbol{w},w_0}{\text{argmin}} \quad \frac{1}{N}\sum_{n=1}^{N}(g(\boldsymbol{x}_n)-y_n)^2 \\ &\approx \underset{\boldsymbol{w},w_0}{\text{argmin}} \quad \mathbb{E}_{\boldsymbol{x},y}\left[(\boldsymbol{w}^T\boldsymbol{x}+w_0-y)^2\right].\end{aligned}$$

Summary of the Result

Theorem (Conditions for Linear Regression = Bayes)

Suppose that all the following three conditions are satisfied:

(i) *The likelihood* $p(\boldsymbol{x}|i)$ *is Gaussian satisfying*

$$p(\boldsymbol{x}|i) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}, \; i \in \{-1,+1\}$$

(ii) *The prior is uniform:* $p_y(+1) = p_y(-1) = \frac{1}{2}$.

(iii) *The number of training samples goes to infinity.*

Then, the linear regression model parameter $(\boldsymbol{w}, w_0)$ *is given by*

$$\boldsymbol{w} = \widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}), \qquad w_0 = -\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_{-1})\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1}),$$

where $\widetilde{\boldsymbol{\Sigma}} \stackrel{def}{=} \boldsymbol{\Sigma}/2$, *and* $\boldsymbol{\Sigma}$ *is the covariance of the Gaussian.*

Sketch of Proof

Let us make some assumptions:

- Likelihood: Gaussian with equal covariance:

$$p(\boldsymbol{x}_n|y=+1) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_{+1})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_{+1})\right\}$$

$$p(\boldsymbol{x}_n|y=-1) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_{-1})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_{-1})\right\}$$

- Prior: Equal prior:

$$p_y(+1) = \frac{1}{2}$$

$$p_y(-1) = \frac{1}{2}.$$

Sketch of Proof

- Taking derivative w.r.t. $(\boldsymbol{w}, w_0)$ yields

$$\frac{d}{d\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x},y}\left[(\boldsymbol{w}^T\boldsymbol{x}+w_0-y)^2\right] = 2\left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{w}+\mathbb{E}[\boldsymbol{x}]w_0-\mathbb{E}[\boldsymbol{x}y]\right)$$

$$\frac{d}{dw_0}\mathbb{E}_{\boldsymbol{x},y}\left[(\boldsymbol{w}^T\boldsymbol{x}+w_0-y)^2\right] = 2\left(\mathbb{E}[\boldsymbol{x}]^T\boldsymbol{w}+w_0-\mathbb{E}[y]\right)$$

- What is $\mathbb{E}[\boldsymbol{x}]$?

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}] &= \mathbb{E}[\boldsymbol{x}|y=1]p_y(+1)+\mathbb{E}[\boldsymbol{x}|y=-1]p_y(-1)\\ &= \boldsymbol{\mu}_1\left(\frac{1}{2}\right)+\boldsymbol{\mu}_{-1}\left(\frac{1}{2}\right) = \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_{-1}).\end{aligned}$$

- What is $\mathbb{E}[\boldsymbol{x}y]$?

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}y] &= \mathbb{E}[\boldsymbol{x}y|y=+1]p_y(+1)+\mathbb{E}[\boldsymbol{x}y|y=-1]p_y(-1)\\ &= (+\boldsymbol{\mu}_1)\left(\frac{1}{2}\right)+(-\boldsymbol{\mu}_{-1})\left(\frac{1}{2}\right) = \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{-1}).\end{aligned}$$

Sketch of Proof

- What is $\mathbb{E}\Big[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T\Big]$?

$$\begin{aligned}
&\mathbb{E}\Big[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T\Big] \\
&= \mathbb{E}\Big[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T | y=+1\Big]\, p_y(+1) \\
&\quad + \mathbb{E}\Big[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T | y=-1\Big]\, p_y(-1) \\
&= \frac{1}{2}\boldsymbol{\Sigma} + \frac{1}{2}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}.
\end{aligned}$$

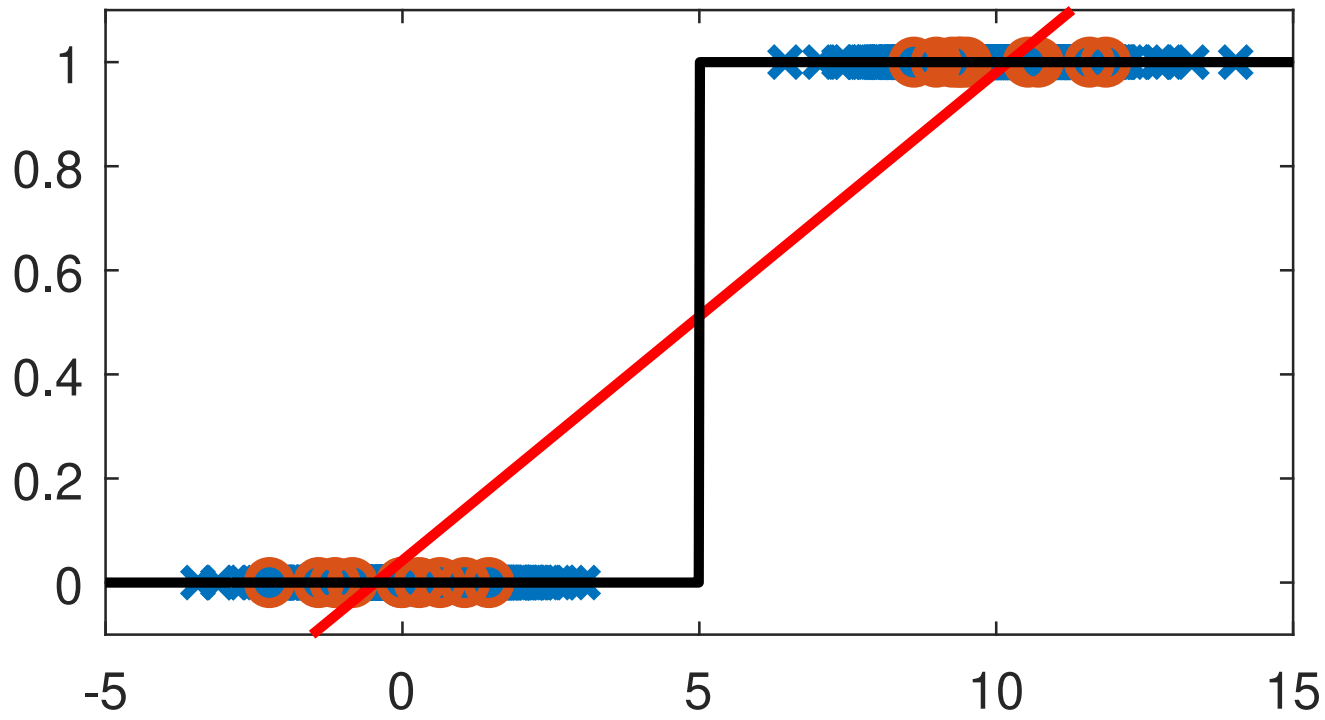
- This will allow us to compute $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$:

$$\mathbb{E}\Big[(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T\Big] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T.$$

- The remaining is just linear algebra. See Appendix.

Implication

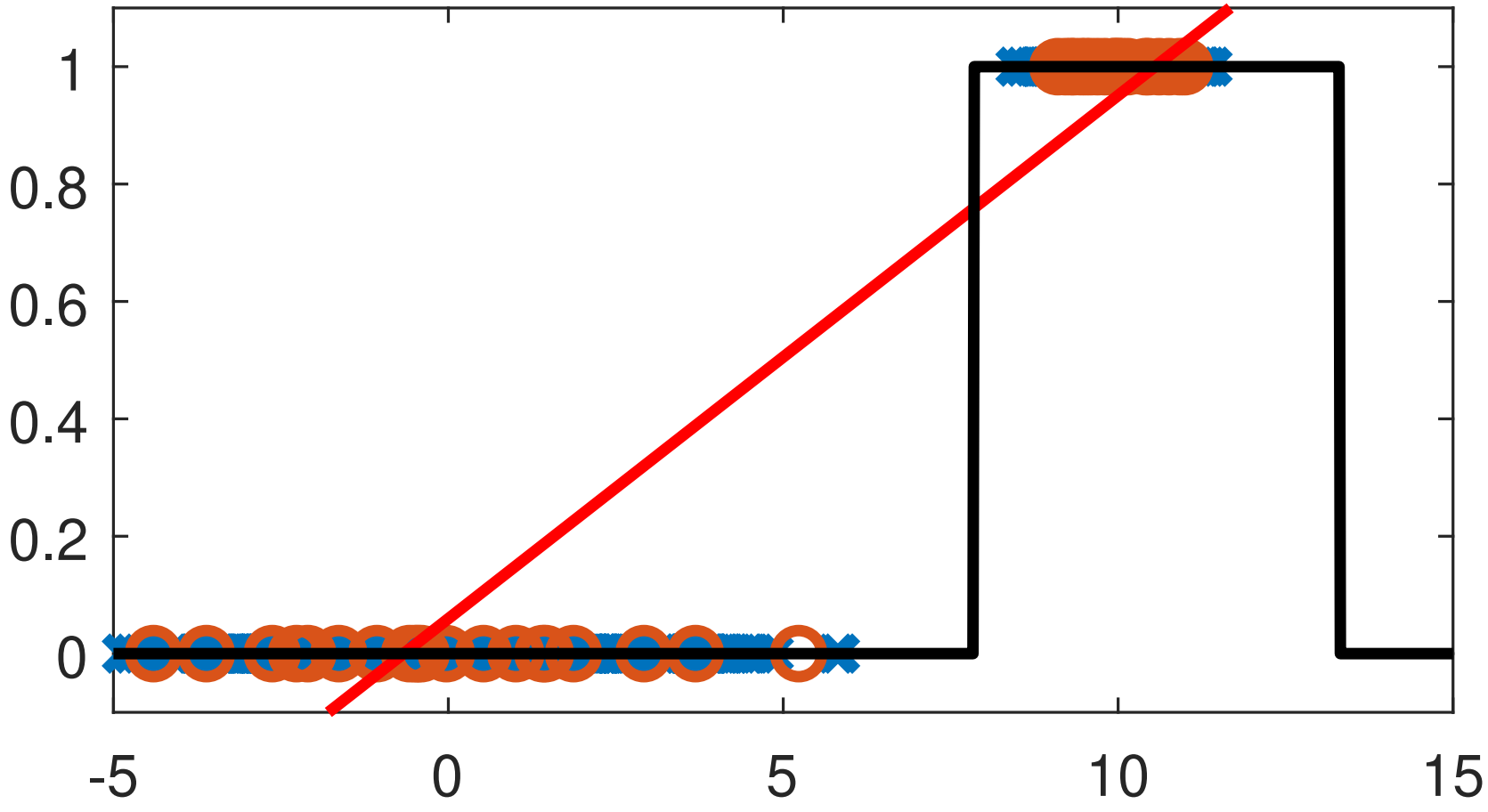
- Linear regression assumes equal covariance for both classes



- Bayesian allows different **variance** $\boldsymbol{\Sigma}_i$.
- They are equal only when number of training samples is large.

When will Linear Regression Go Wrong? (1)

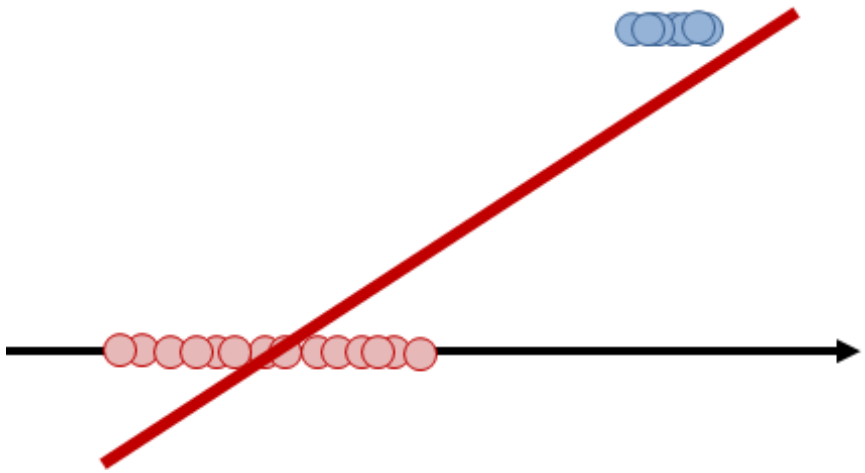
- **Example 1**: When the classes are intrinsically unbalanced.



- Bayesian gives nonlinear decision boundary

When will Linear Regression Go Wrong? (1)

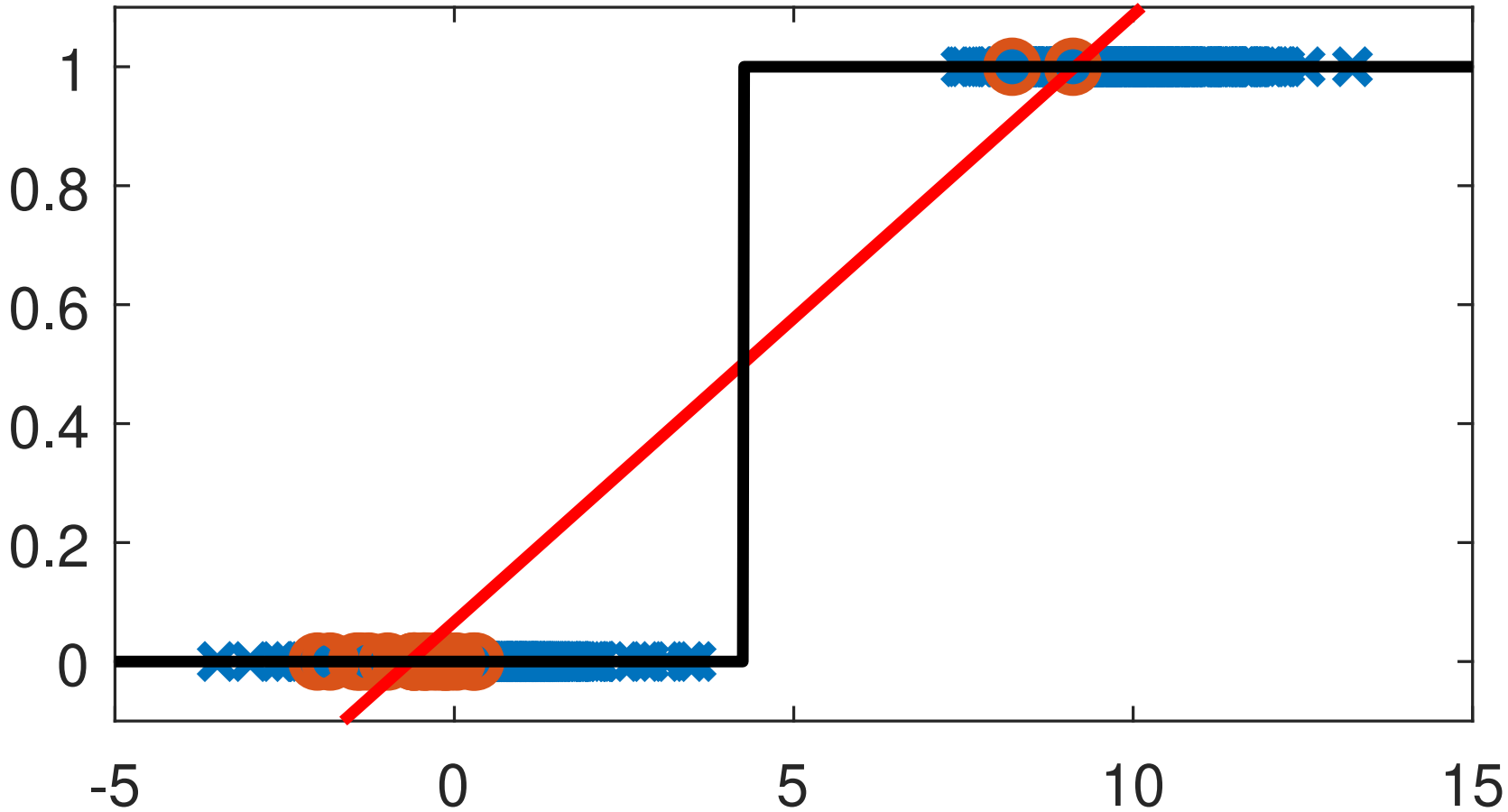
- When the classes are intrinsically unbalanced.
- One class has a significantly larger variance than the other.
- Nothing to do with the number of training samples.
- Regression goes wrong because the big variance class dominates the sum square error.
- So you spend more effort to make that class "happy".



- Bayesian decision rule takes care of this by allowing different $\boldsymbol{\Sigma}_i$.

When will Linear Regression Go Wrong? (2)

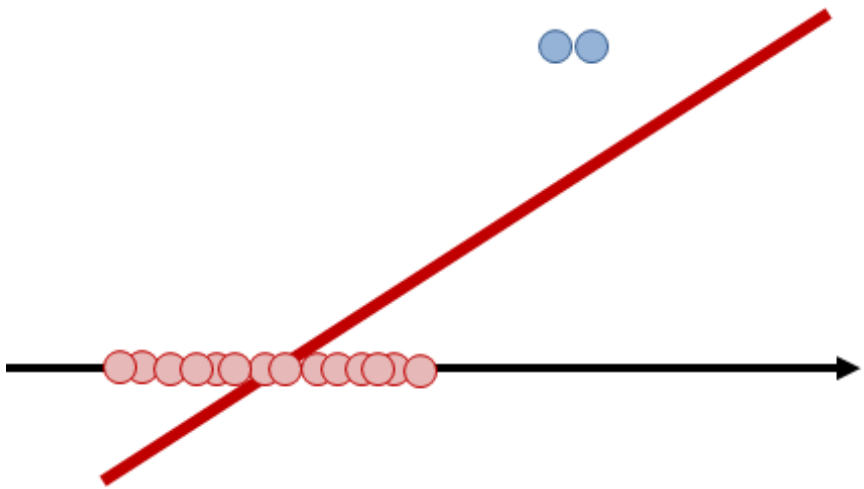
- **Example 2**: When training samples are unbalanced.



- Bayesian performs equally bad.

When will Linear Regression Go Wrong? (2)

- When training samples are unbalanced.
- One class has more training samples than the other class.
- Nothing to do with the intrinsic distribution. You just did not sample the training samples uniformly from the true distribution.
- Regression goes wrong because the more sample class dominate the sum square error.
- So you spend more effort to make the majority "happy".



- Bayesian suffers too because it has a bad estimate of the mean.

Does Regularization Help?

- We can put regularization to linear regression

$$J(\boldsymbol{\theta}) = \|\boldsymbol{A\theta} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

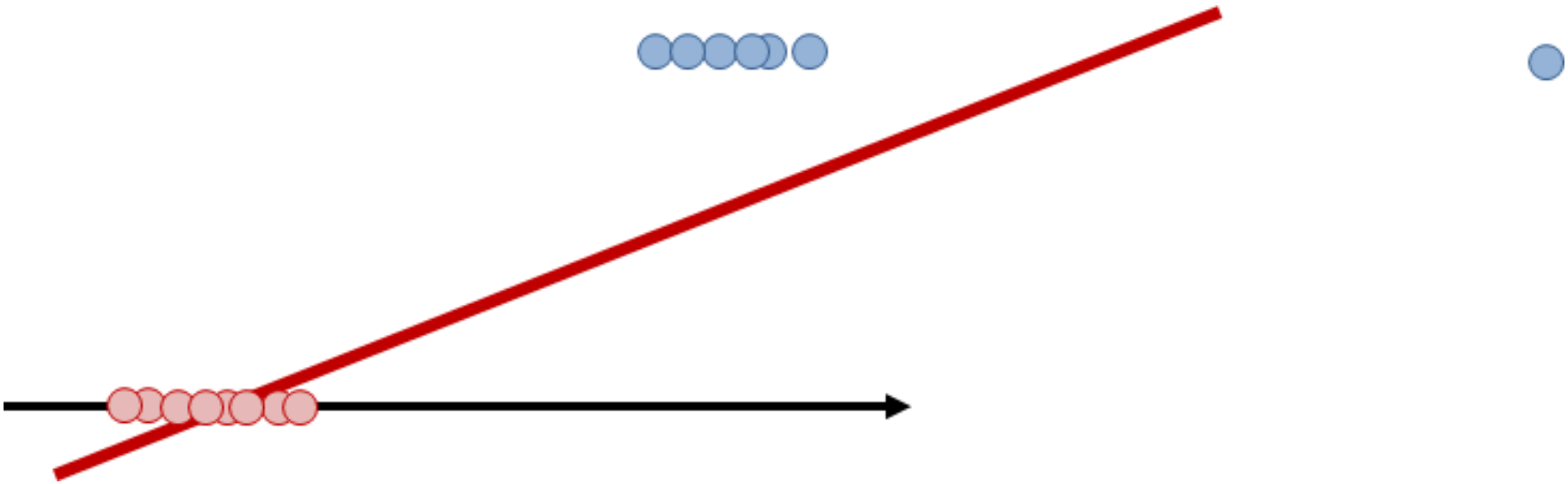
- Can help some bizarre cases when $\boldsymbol{A}$ is rank deficient.
- But what regularization to use? How to control λ?
- Prior in Bayesian is a lot more intuitive.

$$\widehat{\mu} = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}}.$$

- When N is small, we have the prior to control the estimate.
- Linear regression does not have this capability, unless you know what the decision weights should look like.

When will Linear Regression Go Wrong? (3)

- **Example 3**: "Outliers"
- One sample point appears "abnormally"



- Bayesian suffers from the same issue
- But Bayesian can use the prior term to mitigate outliers
- Of course, you can also do data pre-processing in linear regression to remove outliers

Reading List

Linear Regression and Bayesian Decision

- Chris Bishop's *Pattern Recognition*, Chapter 3.1, 4.1
- Hastie-Tibshirani-Friedman's *Elements of Statistical Learning*, Chapter 3.2, 3.4
- Stanford CS 229 Discriminant Algorithms
  http://cs229.stanford.edu/notes/cs229-notes1.pdf

Appendix

Proof of Main Result

By following the steps in the proof sketch, we have shown that

$$\frac{d}{d\boldsymbol{w}}\mathbb{E}_{\boldsymbol{x},y}\left[(\boldsymbol{w}^T\boldsymbol{x}+w_0-y)^2\right] = 2\left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{w}+\mathbb{E}[\boldsymbol{x}]w_0-\mathbb{E}[\boldsymbol{x}y]\right)=0$$
$$\frac{d}{dw_0}\mathbb{E}_{\boldsymbol{x},y}\left[(\boldsymbol{w}^T\boldsymbol{x}+w_0-y)^2\right] = 2\left(\mathbb{E}[\boldsymbol{x}]^T\boldsymbol{w}+w_0-\mathbb{E}[y]\right)=0$$

- Look at the second equation

$$\begin{array}{llll} -\mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T\boldsymbol{w} & -\mathbb{E}[\boldsymbol{x}]w_0 & +\mathbb{E}[\boldsymbol{x}]\mathbb{E}[y] & =0 \\ +\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{w} & +\mathbb{E}[\boldsymbol{x}]w_0 & -\mathbb{E}[\boldsymbol{x}y] & =0 \end{array}$$

- This gives us

$$(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]-\mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T)\boldsymbol{w}+0-(\mathbb{E}[\boldsymbol{x}y]-\mathbb{E}[\boldsymbol{x}]\mathbb{E}[y])=0.$$

Proof of Main Result

- Therefore, we have

$$\underbrace{(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T)}_{\boldsymbol{\Sigma}}\boldsymbol{w} + 0 - \Big(\underbrace{\mathbb{E}[\boldsymbol{x}y]}_{=\frac{1}{2}(\boldsymbol{\mu}_{+1}-\boldsymbol{\mu}_{-1})} - \mathbb{E}[\boldsymbol{x}]\underbrace{\mathbb{E}[y]}_{=0}\Big) = 0.$$

- This means that

$$\boldsymbol{\Sigma}\boldsymbol{w} = \frac{1}{2}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}),$$

- which gives us

$$\boldsymbol{w} = \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}).$$

- Compare to the Bayesian decision rule for equal covariance:

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}).$$

- The only difference is the factor $1/2$.

Proof of Main Result

- Now let us determine w_0.
- Look at the second equation again:

$$\mathbb{E}[\boldsymbol{x}]^T \boldsymbol{w} + w_0 - \mathbb{E}[y] = 0$$

- This means

$$\begin{aligned} w_0 &= \mathbb{E}[y] - \mathbb{E}[\boldsymbol{x}]^T \boldsymbol{w} \\ &= 0 - \left(\frac{1}{2}(\boldsymbol{\mu}_{+1} + \boldsymbol{\mu}_{-1})\right)^T \boldsymbol{w} \\ &= 0 - \left(\frac{1}{2}(\boldsymbol{\mu}_{+1} + \boldsymbol{\mu}_{-1})\right)^T \left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})\right) \\ &= -\frac{1}{4}(\boldsymbol{\mu}_{+1} + \boldsymbol{\mu}_{-1})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}). \end{aligned}$$

Proof of Main Result

- If we want to write the decision boundary as $\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) = 0$,
- then we can show that

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}_0) = \left(\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})\right)(\boldsymbol{x} - \boldsymbol{x}_0)\,.$$

- Since

$$w_0 = -\frac{1}{4}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} + \boldsymbol{\mu}_{-1}),$$

- in order to make $w_0 = \boldsymbol{w}^T\boldsymbol{x}_0$, we should choose

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_{+1} + \boldsymbol{\mu}_{-1}).$$

- This is the same as the Bayesian decision rule with equal covariance.