# ECE595 / STAT598: Machine Learning I
## Lecture 13.2: Connection between Linear Regression and Bayesian

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

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

## Sketch of Proof

Let us make some assumptions:

- Likelihood: Gaussian with equal covariance:

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

- Prior: Equal prior:

$$
\begin{aligned}
p_y(+1) &= \frac{1}{2}\\
p_y(-1) &= \frac{1}{2}.
\end{aligned}
$$

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

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

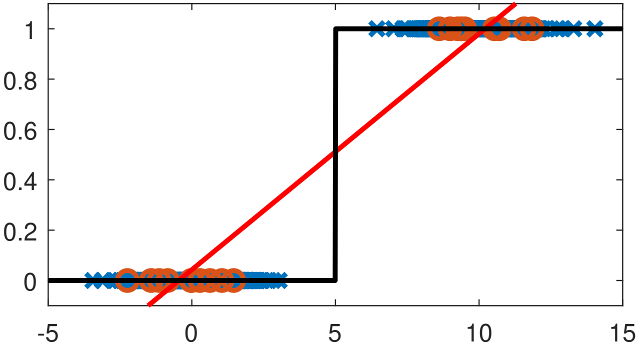
- This will allow us to compute $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$:

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

- The remaining is just linear algebra. See Appendix.

# Implication

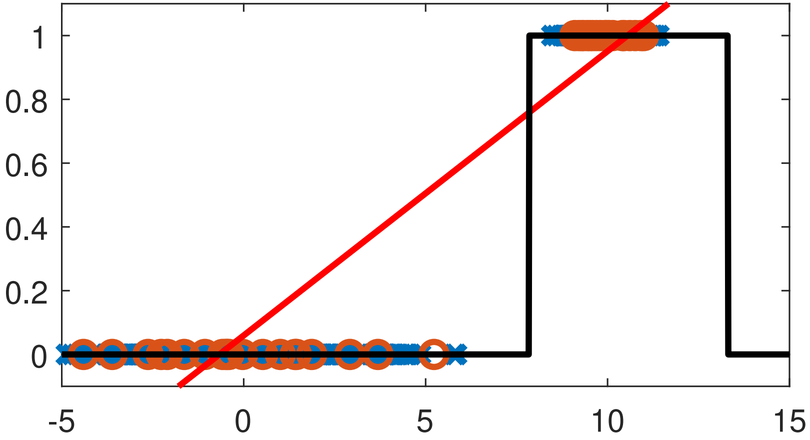
- Linear regression assumes equal covariance for both classes




- 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.

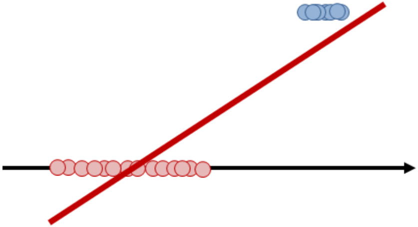



-

- Bayesian gives nonlinear decision boundary

# When will Linear Regression Go Wrong? (1)

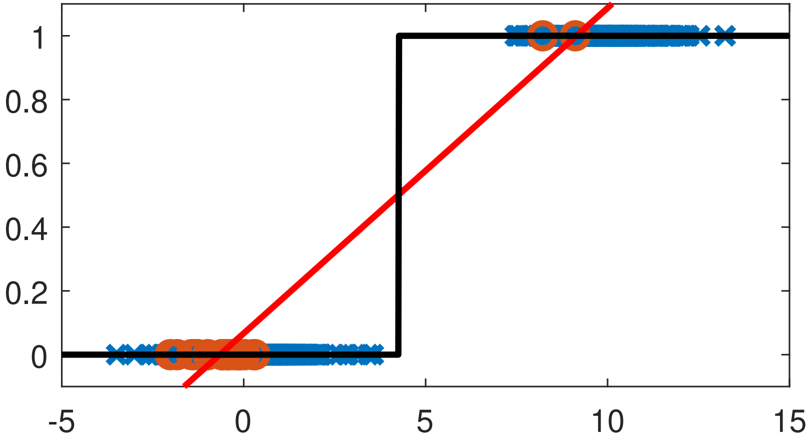
- When the classes are intrinsically unbalanced.
- One class has a significantly larger variance than the other.
- Nothing to do with the number of training samples.
- Regression goes wrong because the big variance class dominates the sum square error.
- So you spend more effort to make that class "happy".



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

# When will Linear Regression Go Wrong? (2)

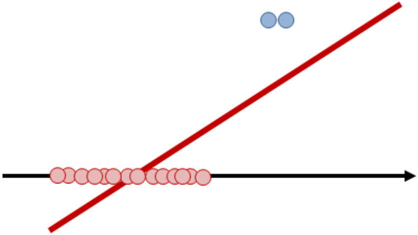
- **Example 2**: When training samples are unbalanced.




- 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

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

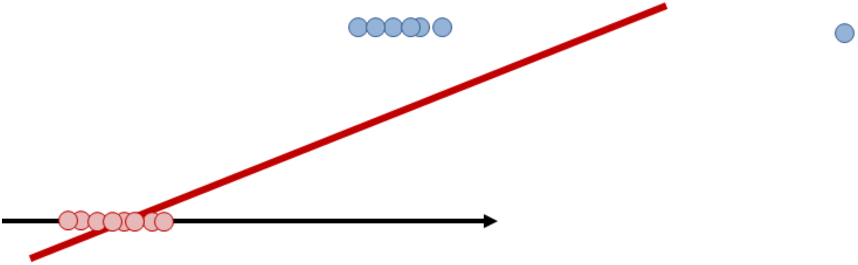
- Can help some bizarre cases when $\boldsymbol{A}$ is rank deficient.
- But what regularization to use? How to control $\lambda$?
- Prior in Bayesian is a lot more intuitive.

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

- 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

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

- 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

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

- 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.