DASC4113 Machine Learning

Ukash Nakarmi

Lecture 4

CLASSIFICATION

Learning Objectives

In this class, we will learn about following concepts:

Logistic Regression

- Given some input features how do we classify whether the features belong to class A or class B ?
- How do we formulate classification problems when numbers of output classes is more than 2?

Generative Models for Classification (GMC)

- What are other tools besides logistic regression for classification?
- What is the primary difference between logistic regression and GMC?

Classification: Example Problems

1. Given some predictors such as income and balance of an individual, a person would default the credit or not?
2. A patient comes with some symptoms, what is the probability the individual has medical condition A, B or C?
3. Predict an Online banking transaction is fraudulent or not?

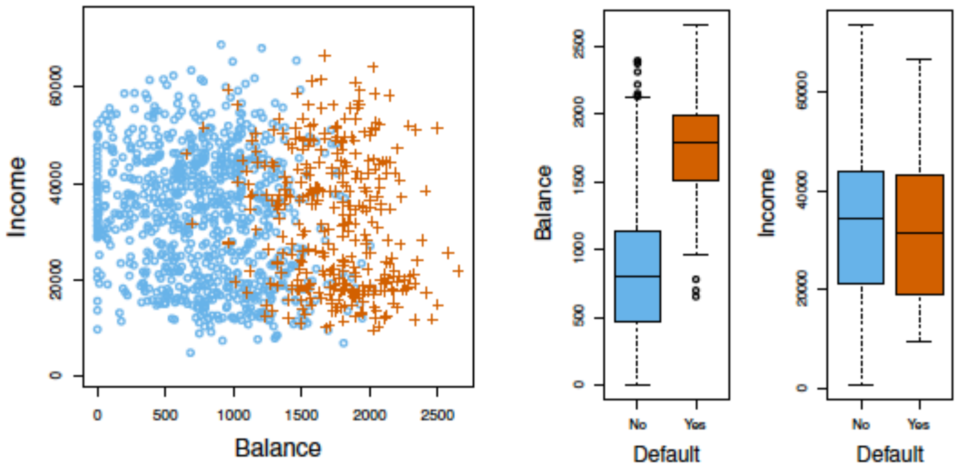


Fig: 4.1 The 'Default' dataset

Classification: Examples

1. Given some predictors such as income, balance, of an individual, a person would default the credit or not?
2. A patient with some symptoms, what is the probability, the individual has medical condition A, B or C?
3. Predict an Online banking transaction is fraudulent or not?

On all 3 examples above, we some characteristics about the response

Response:

- Categorical
- Qualitative
- Involves probability

Logistic Regression

Let's consider the Credit Default Problem:

Given the balance what is the probability the credit will be defaulted

$$\Pr(default = Yes|balance)$$

$$\Pr(Y|X)$$

Q. What happens if we try to fit using Linear Regression?

$$p(X) = \beta_0 + \beta_1 X.$$

- The right-hand side of equation is not bounded to 0-1
- Linear regression using OLS enforces natural ordering, which is not true for classification task.

Example: Natural Ordering

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure;} \\ 2 & \text{if } \texttt{stroke;} \\ 3 & \text{if } \texttt{drug overdose.} \end{cases}$$

Logistic Regression

We model a function p(X) = Pr(Y|X) as a function that gives value between 0 -1

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

Step 1: Functional form of the model
Recall: 2 steps in finding the function f()
Note: p(X) is not linear on X anymore

The function is termed Logistic Function/(Sigmoid)

We can re-write it as:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}.$$

Odds

- No new information but more interpretable.
- Ratio of probability of something happening to probability of not happening
- Maps probability between 0 -1 to 0 - ∞

Logistic Regression

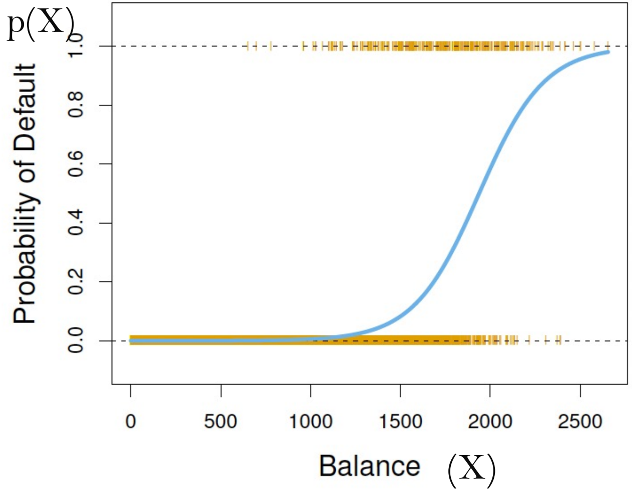
Logit: $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Log of Odds (Logit)

Recall: p(X) is not linear on X in logistic regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}.$$

With some Manipulation, we made Logit as linear in X.



- Logit form makes it easier to interpret β_1.
- One unit change in X increases the Logit by β_1 (Note, p(X) is not linear in X).
- Slope is not constant (for X vs P(X)).

Logistic Regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Step 2: Estimating Parameters β

Recall: In Linear Regression we estimated β by minimizing least squares between the predicted and the true value.

In Logistic Regression, we use *likelihood* function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

Estimate β such that the likelihood function is maximized.

Logistic Regression

The Likelihood Function

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

Estimate β such that the likelihood function is maximized.

Multiple Logistic Regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Recall: Single input case

More than 1 input variable

Step1. Logit takes the form:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

X_1, X_2 … X_p are p input variable predictors

Step 2. Estimate β by maximizing likelihood function.

$$l(\beta_0, \beta_1, \dots \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

Simple and Multiple Logistic Regression : Comparison

Dataset Preview: 'Default'

default	student	balance	income	default2	student2
No	No	729.526495	44361.625074	0	0
No	Yes	817.180407	12106.134700	0	1
No	No	1073.549164	31767.138947	0	0
No	No	529.250605	35704.493935	0	0
No	No	785.655883	38463.495879	0	0

Simple and Multiple Logistic Regression : Comparison

Simple Logistic Regression Coefficients : Only student status as Input

	Coefficient	Std. error	z-statistic	p-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Multiple Logistic Regression Coefficient using balance, income and student status as Inputs

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Simple and Multiple Logistic Regression : Comparison

SLR: Student Coeff: 0.4049

- +ve Coeff
- The probability someone defaults the credit card is higher if Student

Not student

Are student

MLS: Student Coeff: - 0.6468

- - ve Coeff
- The probability someone defaults the credit card is lower if Student

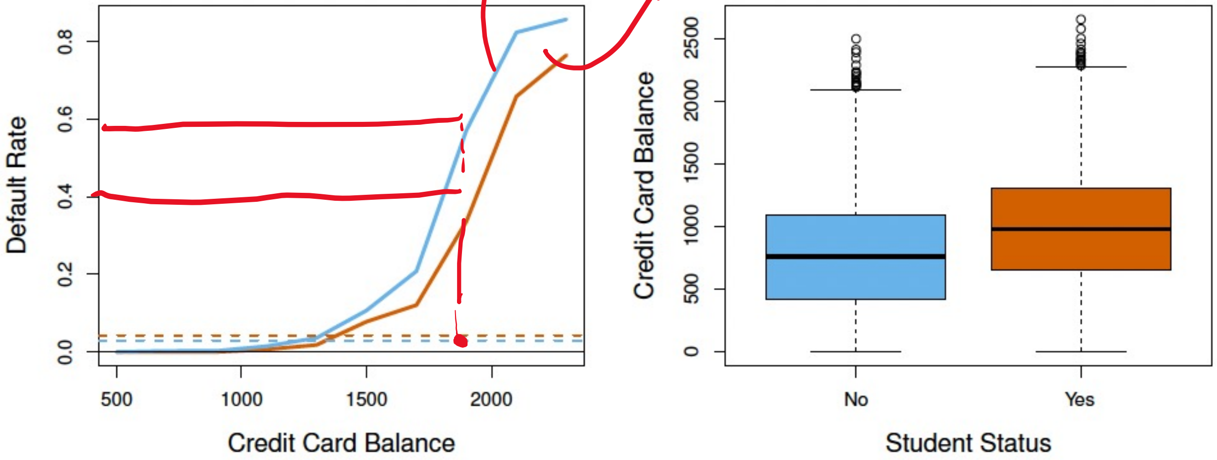


Fig: 4.3

Why we have different stories in SLR vs MLS ?

- More dependent on Balance than student status.
- For the same value of balance, students are less likely to default the credit.

Multinomial Logistic Regression

More than 2 output categories.

A. Baseline Approach:

- Model it as two categories: one category as a baseline class and the rest all categories as another class.
- Assume class K is the baseline

$$Y = \begin{cases} 1 & \text{if epileptic seizure;} \\ 2 & \text{if stroke;} \\ 3 & \text{if drug overdose.} \end{cases}$$

The functional form takes the form:

Class Input variable

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$\forall\, k = 1, 2, \ldots, k-1$

and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Multinomial Logistic Regression

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p.$$

The log odds between any pair of classes is linear in the features

- The decision to treat the Kth class as the baseline is unimportant
- No matter which class we use as the base class,
- The values of coefficients may be different but,
- The log odds (logits) and prediction remains the same.

Multinomial Logistic Regression

B. Softmax Function

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Recall: Baseline Approach

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$\forall\, k = 1, 2, \ldots, k-1$

- Rather than estimating coefficients for $K-1$ classes separately, we estimate coefficients for all K classes.

Log Odds between two classes k and k':

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p.$$

Generative Model for Classification

We already have Logistic Regression:

Q. Why do we need anything else?

Q. How is Generative Model different than Logistic Regression?

- Naturally transitions from two classes to more than two classes.
- Considers the knowledge about the distribution (probability density function) of data in each class.

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naïve Bayes

Bayes' Theorem

Revisit: Bayes' Theorem

What it is?
Why is it important ?

- A, B — Events
- P(A), P(B) — Independent Probabilities of A and B respectively.
- P(A|B) — Conditional Probability of A given B is true
- P(B|A) — ~

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

- Allows us to update our predictions based on new information (conditional probabilities)

Examples:

- What is the probability the Amazon co. price falls given the dow jones average falls?
- What is the probability a patient has a cancer given the test result is positive?

Revisit: Bayes' Theorem

Cancer Test Example: Scenario

- A person takes a cancer test and gets the positive result.
- But we have some information about the cancer statistics in general and test efficiency.

  - Event A: P(A)
    Probability of having cancer (Probability any person has a cancer)
  - Event B : P(B)
    Probability the test is positive
    (Probability of Having any test positive, true positive + false positive)
  - Test Accuracy: P(B/A)
    Probability the test is positive given the person has cancer.
    (Probability of True Positive)

Our interest:

P(A/B) : Probability a person has a cancer given the test is positive.

Revisit: Bayes' Theorem

Cancer Diagnostic Test Example:

- P(A) = 0.1
- P(B) = 0.86

(Probability of Having any test positive, true positive + false positive)

- P(B/A) = 0.8

True Positive

Our interest:

P(A/B) : Probability a person has a cancer given the test is positive.

$$P(A/B) = \frac{P(A)P(B/A)}{P(B)} = \frac{0.1*0.8}{0.86} = 0.093$$

Q. What happens to P(A/B) when false positive = 0

Generative Models for Classification

- Considers the knowledge about the distribution (probability density function) of data in each class

Bayes' Theorem: Different flavor

Alternate Notaion: $p_k(x)$

$$\underbrace{\Pr(Y = k}_{\text{Observation belongs to class k}}|\underbrace{X = x}_{\text{Observation}}) = \frac{\overbrace{\pi_k}^{\text{Probability of class k}} \overbrace{f_k(x)}^{\Pr(X \mid y = k) \text{ (Density function)}}}{\underbrace{\sum_{l=1}^{K} \pi_l f_l(x)}_{\text{Total Probability}}}$$

Try compare with original form:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Generative Models for Classification

Alternate Notaion: $p_k(x)$

Probability of class k

$$\Pr(\underbrace{Y = k}_{\text{Observation belongs to class k}} \mid \underbrace{X = x}_{\text{Observation}}) = \frac{\pi_k \overbrace{f_k(x)}^{\Pr(X \mid\ y\, =\, k)\ \text{(Density function)}}}{\underbrace{\sum_{l=1}^{K} \pi_l f_l(x)}_{\text{Total Probability}}}$$

Unknowns:

π_k : Can be calculated using the sample data.

$f_k(x)$: Not trivial to know exactly. We make assumptions about the distribution of data

Generative Models for Classification

Based on the assumptions we make on density function (**distribution**):

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naïve Bayes

We will see how some simple assumptions about $f_k(x)$ changes the way we compute:

$\Pr(y = k \mid X)$

$Alternative\ Notation$: $p_k(x)$

Linear Discriminant Analysis (LDA)

Case 1: p = 1, i.e. We have only one input variable

Alternate Notaion: $p_k(x)$

$$\Pr(Y=k|X=x)=\frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

Recall: Our Goal is:

- We want to estimate $f_k(x)$
- Compute $p_k(x)$
- Assign/Classify our observation x to the class which has highest $p_k(x)$

UNIVERSITY OF ARKANSAS

Linear Discriminant Analysis (LDA)

Case 1: p = 1, i.e. We have only one input variable

If we assume, $f_k(x)$ is normal/Gaussian, and only one input variable:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

μ_k: Mean of class k
σ_k^2: Variance of class k

Further Assume : $\underline{\sigma_1^2 = \sigma_2^2 = \cdots \sigma_K^2 = \sigma^2}$ i.e. All classes have same variance.

$Y \Rightarrow$ $P \Rightarrow$ Size of a Tumor.

Chronic | Cancer | No Cancer

μ_{Chro} μ_{Cancer} $\mu_{nocancer}$

UNIVERSITY OF ARKANSAS

Linear Discriminant Analysis (LDA)

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Case 1: p = 1, i.e. We have only one input variable

Then our probability $p_k(x)$ expression takes the form:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}.$$

Taking log:

$$\boxed{\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)}$$

We assign our observation (data) to class k for which $\delta_k(x)$ is maximum

Linear Discriminant Analysis (LDA)

Case 1: p = 1, i.e. We have only one input variable

$$\boxed{\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)}$$

Special Case:

When $K = 2$, Only two classes (Binary Classification)

$\pi_1 = \pi_2$ i.e. Both classes have same probabilities. Observation for each class are equally likely

Our Decision Rule becomes :

If: $\boxed{2x\left(\mu_1 - \mu_2\right) > \mu_1^2 - \mu_2^2,}$

Assign to class k = 1

Else:

Assign to Class k = 2.

Our Decision Boundary becomes :

Point where $\delta_1(x) = \delta_2(x)$, *i.e. x could be in any class.*

$$\boxed{x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.}$$

Linear Discriminant Analysis

Example:

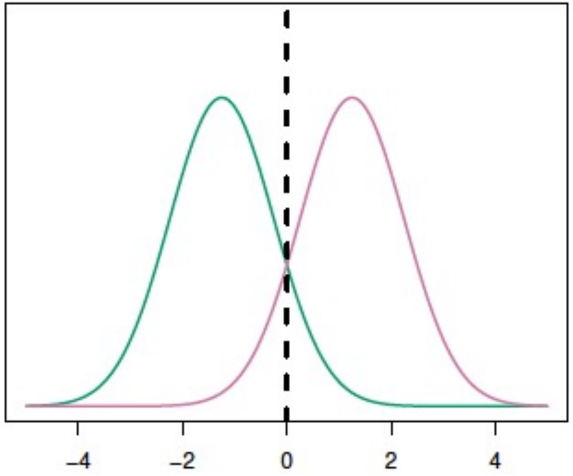
X is 1 dimensional, i.e only one input variable

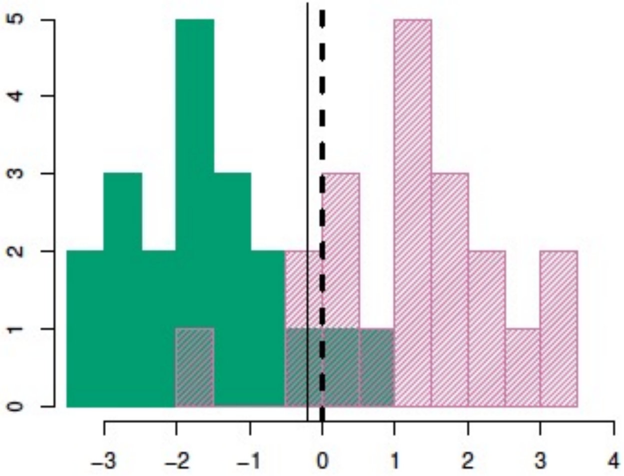
μ_1 = -1.25, μ_2 = 1.25

$\pi_1 = \pi_2$ = 0.5

K =2

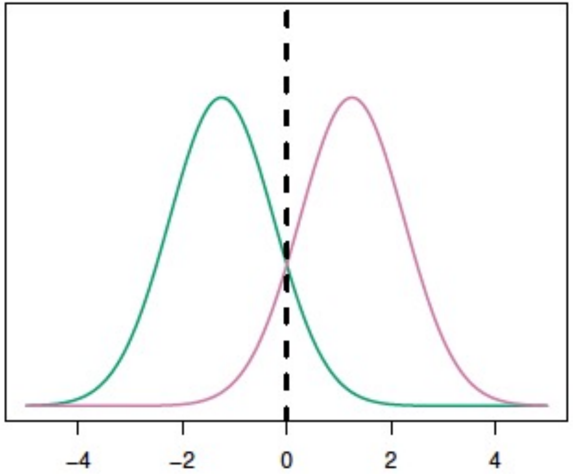
Fig 4.4

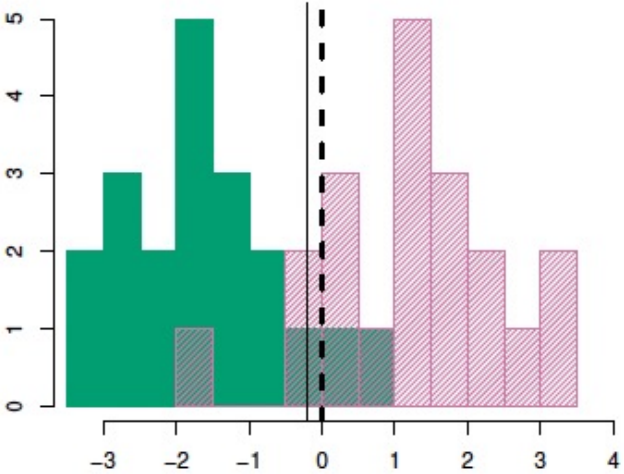




Decision boundary is at x = 0. (Doesn't always have to be so.)

Linear Discriminant Analysis





$$\boxed{\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)}$$

Q. How do we actually compute each of the term in this equation?

Linear Discriminant Analysis

Compute from the sample Data we have:

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

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k/n.$$

$$\boxed{\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)}$$

Q. Why do we call this method Linear Discriminant Analysis?

Linear Discriminant Analysis

Case 2: p >1, i.e. We have more than 1 input variable. Multivariate Inputs

$$X = (X_1, X_2, \ldots, X_p)$$ p Input Variables

Assumptions: Multivariate Normal/Gaussian

- Each predictor follow 1 D Normal Distribution.
- Some correlation between each predictor pairs
- So, our predictors are drawn from multivariate Gaussian/Normal Distribution $X \sim N(\boldsymbol{\mu}, \Sigma)$

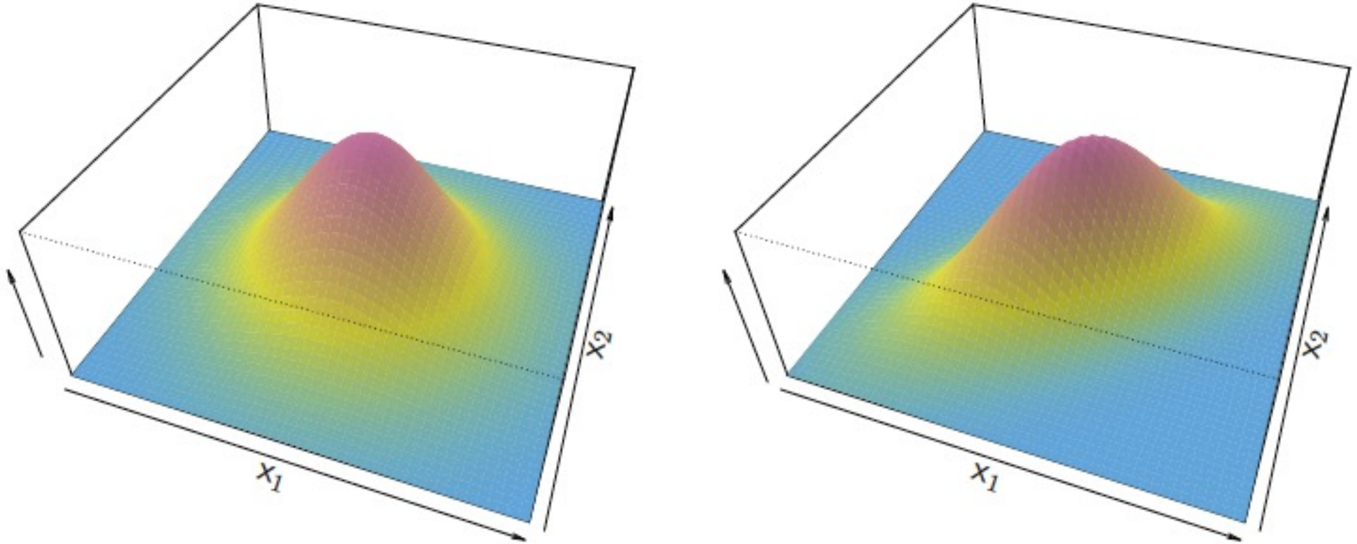
$\boldsymbol{\mu}$: p length vector with elements being mean of each predictor
Σ : Cov (X) is a $p \times p$ covariance matrix. Every class has same covariance.

Let's take an example of p = 2

Linear Discriminant Analysis

Let's take an example of p = 2

Fig: 4.5



Two multivariate Gaussian Distribution with p = 2.

Left: $Var(X_1) = Var(X_2), Cor(X_1, X_2) = 0;$

Right: Correlated predictors and have different variance.

Linear Discriminant Analysis

Case 2: p >1, i.e. We have more than 1 input variable. Multivariate Inputs

$p \times p$

$X = (X_1, X_2, \ldots, X_p)$ p Input Variables

Our predictors are drawn from multivariate Gaussian/Normal Distribution $X \sim N(\boldsymbol{\mu}, \Sigma)$

Multivariate Gaussian Density:

$$\boxed{f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}$$

Decision Rule: Assign observation x to class k for which $\delta_k(x)$ is maximum

$$\boxed{\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k}$$

Linear Discriminant Analysis

Decision Rule: Assign observation x to class k for which $\delta_k(x)$ is maximum

$$\boxed{\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k}$$

Each class have individual p length mean

But share same co-variance matrix

Note: This is what will change in QDA

All notations have similar meaning as in single input case (p = 1)

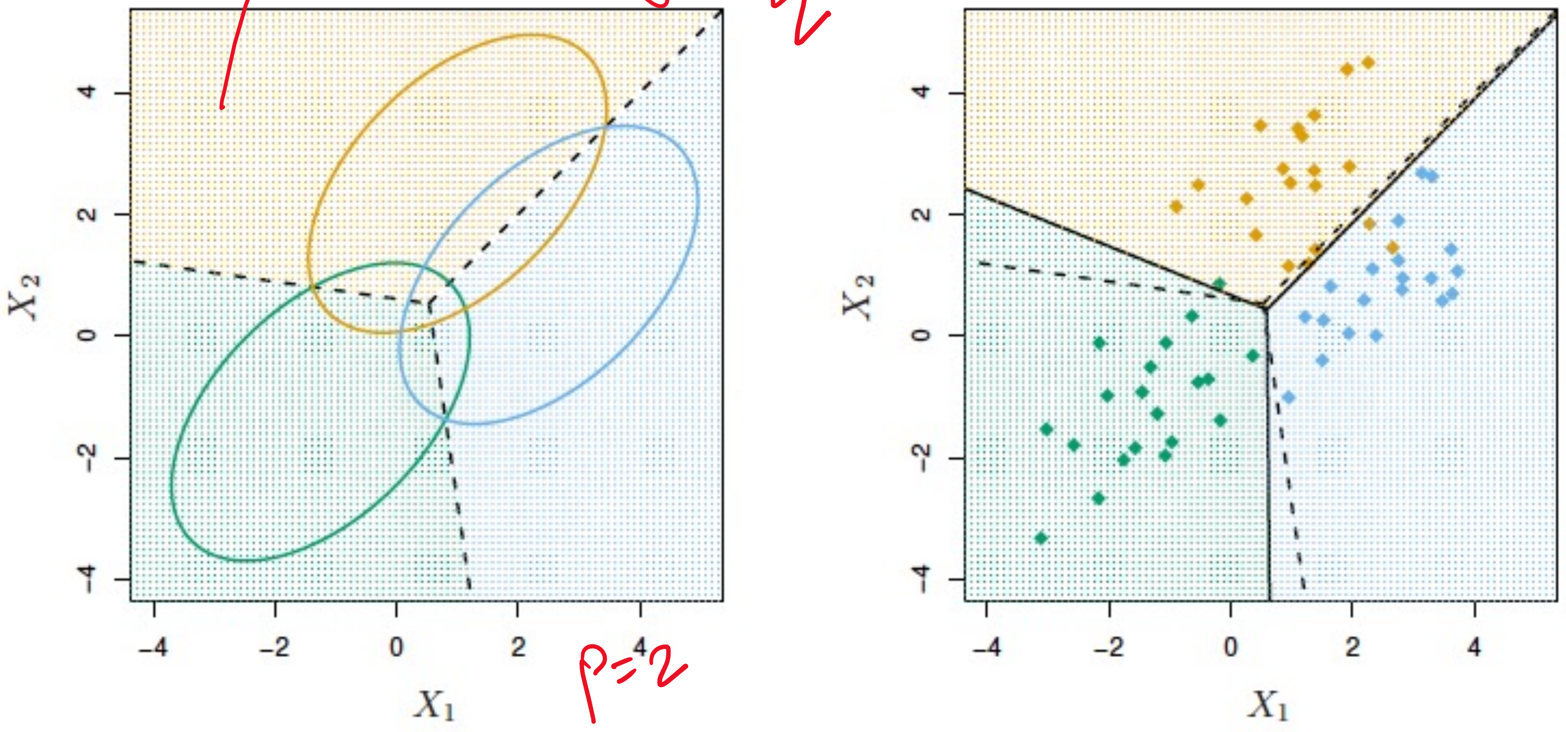
Decision Boundary: Values of x where probability are same, i.e. $\delta_k(x) = \delta_l(x)$

If we assume each class has same probability, Decision Boundary can be calculated using:

$$x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k = x^T\Sigma^{-1}\mu_l - \frac{1}{2}\mu_l^T\Sigma^{-1}\mu_l$$

Linear Discriminant Analysis

μ_k: π_k K=3 Σ



p=2

p = 2, K = 3. Left: True data and Distribution. Right: 20 LDA using 20 samples from each class.

Quadratic Discriminant Analysis

Recall: In LDA, assumptions were:

- Each Class are drawn from a multi-variate Gaussian Distribution
- Each class have different mean ($\boldsymbol{\mu}_k$)
- Each class share same co-variance matrix ($\boldsymbol{\Sigma}$)

- In QDA,
- Each class has its own covariance ($\boldsymbol{\Sigma}_k$)

- Hence, The Gaussian Density for observation in kth class is:

$$X \sim N(\mu_k, \Sigma_k)$$

Quadratic Discriminant Analysis

Decision Rule:

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k \\ &= -\frac{1}{2}x^T \boldsymbol{\Sigma}_k^{-1} x + x^T \boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log \pi_k\end{aligned}$$

Q. Why it is called Quadratic Discriminant?

Naïve Bayes'

$k = 1, 2, 3$

k = Chronic

p = 3

Tumor size

Age

Sex

X

LDA: Each class has different mean, same covariance.
QDA: Each class has different mean, different covariance.

Naïve Bayes' makes different assumption:

- In each class k, input variables are independent to each other. (p predictors are independent)

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

Q. What does this assumption do to the $\boldsymbol{\Sigma}_k$?/How would $\boldsymbol{\Sigma}_k$ change with this assumption?

$\boldsymbol{\Sigma}_k$ becomes Diagonal Matrix

Q. How do we estimate $f_{kj}(x)$

Chronic.

$P = 3$

Observation	Tumor Size	Age	Sex
1	5	23	M
2	6	72	F
3	20	65	F
4	32	7	F
5	70	52	M

$f_{chronic}(x)$

$f_{chronic}^{Tumor}(x)$

$f_{chronic}^{age}(x)$

$f_{chronic}^{sex}(x)$

$f_{chronic}(x)$

Naïve Bayes'

- In each class k, input variables are independent to each other. (p predictors are independent)

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

Q. How do we estimate $f_{kj}(x)$

Note: Naïve Bayes not necessarily assume $f_{kj}(x)$ to be Gaussian always.

1. If we assume them to be Gaussian, we can estimate μ_{kj} *and* σ_{kj}^2 for each class and predictor using training data.
2. If we do not make gaussian assumption:

    Plot histogram, smooth the histogram and learn the distribution.

Naïve Bayes'

Assumption:

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

Decision Rule

$$\boxed{\Pr(Y = k|X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \cdots \times f_{lp}(x_p)}}$$

LDA, QDA and Naïve Bayes' Comparison

Recall: Our Goal was to find (evaluate) this expression:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

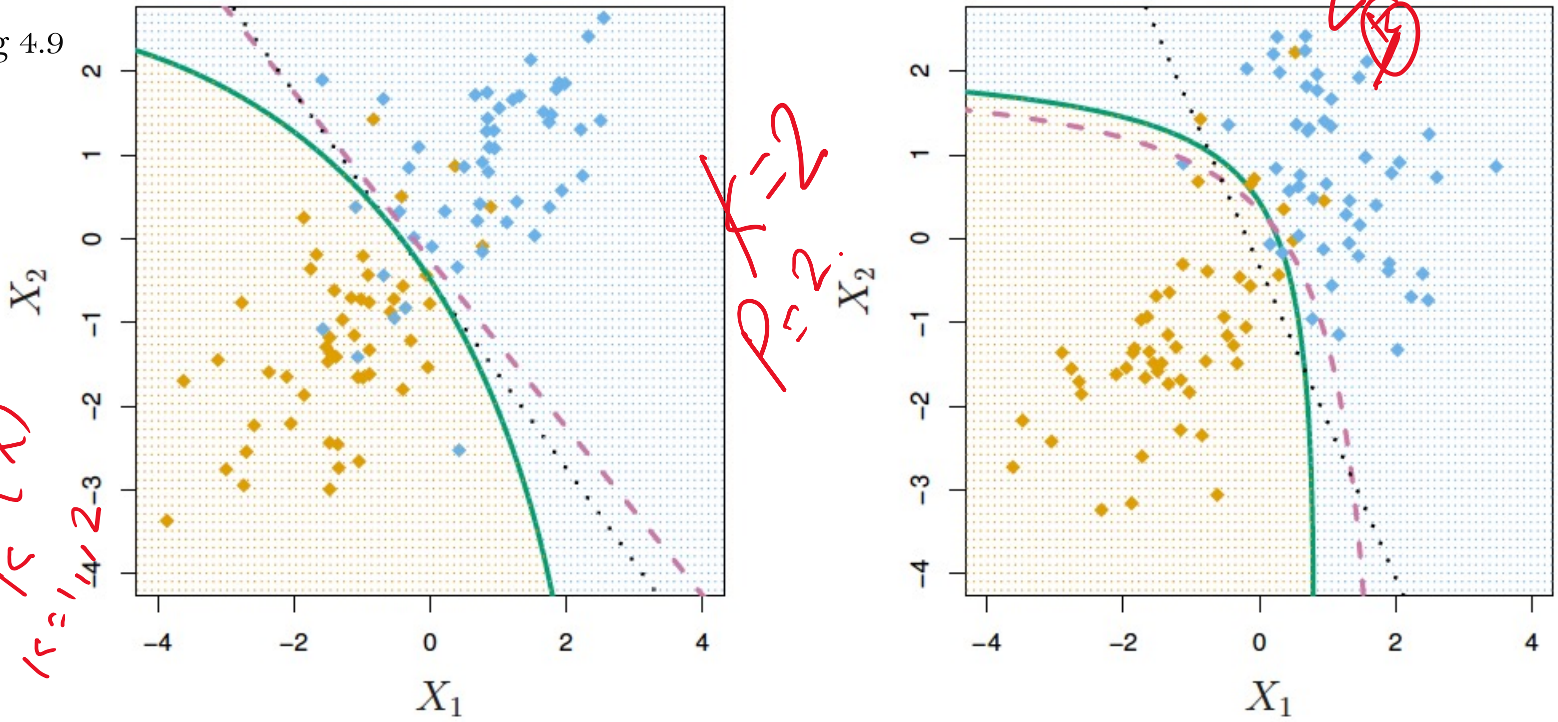
Hope we can appreciate how different assumptions changes the way we estimate $f_k(x)$.

Choice depends on:

- Some are easier compared to other.
- Depends on the amount of data we have.
- Depends on how much we know the system.
- Which assumption best fit our system

LDA, QDA and Naïve Bayes' Comparison

Fig 4.9



Q. Given these two data sets (Left and Right), in which case you would make assumption Σ is same for each class?

Q. Which one would you use LDA, QDA or Naïve Bayes in each case?

Generalized Linear Regression

A Bike Share Example:

Reading Assignment 3

Book: ISLR Section 4.6

Generalized Linear Regression

Data Type of Target Variable (Y) Perspective:

- Case 1: Continuous(Quantitative) -> Regression
- Case 2: Categorical (Qualitative) -> Classification

Case 3: Y is neither Quantitative nor Qualitative

Bike Share Data: Y is neither Quantitative nor Qualitative

Y ~ Number of Hourly Users/Bikers in bikeshare program

Non-negative integer values (counts)

Generalized Linear Regression

Bike Share Data: Data Preview

Generalized Linear Regression

Bike Share Data:

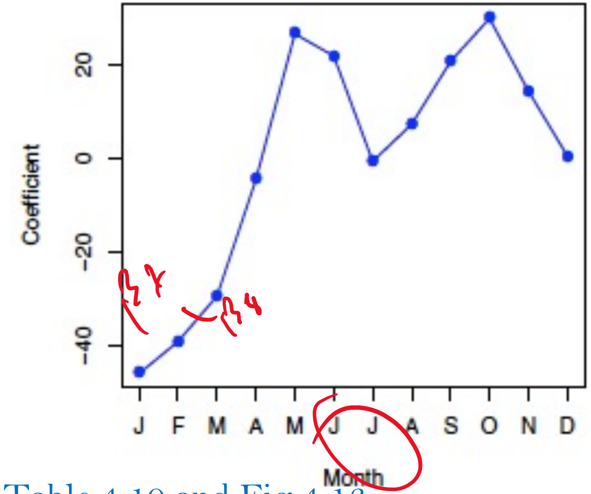
Input(Notation)	Type	Options
Month of the Year (mnth)	Categorical	1 - 12
Hour of the day (hr)	Categorical	0 - 23
Working Day (workingday)	Categorical	1, if: Neither Weekend nor Holidays, else 0
Temperature (temp)	Cont…	
Weather Situation (weathersit)	Categorical	Clear (Baseline) vs: misty/cloudy light rain/ ligh snow heavy rain /heavy snow

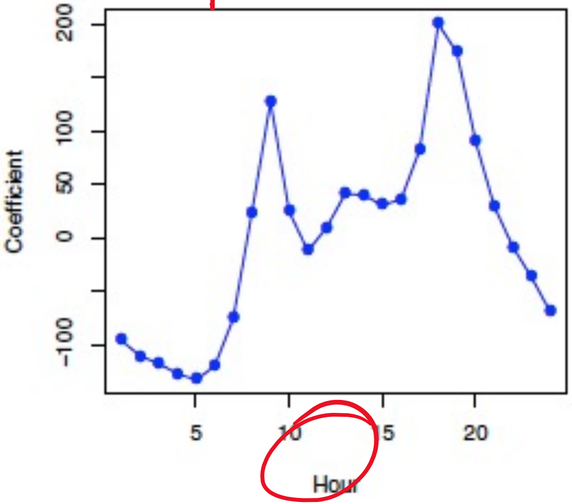
Output : Number of bikers per hour (Counts)

Generalized Linear Regression

Bike Share Data: What do we get if we do Linear Regression ?

		Coefficient	Std. error	z-statistic	p-value
β_0	Intercept	73.60	5.13	14.34	0.00
	workingday	1.27	1.78	0.71	0.48
β_1	temp	157.21	10.26	15.32	0.00
β_2	weathersit[cloudy/misty]	-12.89	1.96	-6.56	0.00
	weathersit[light rain/snow]	-66.49	2.97	-22.43	0.00
	weathersit[heavy rain/snow]	-109.75	76.67	-1.43	0.15





Q. Why Linear Regression is not a good model for this data ?

Table 4.10 and Fig 4.13

Generalized Linear Regression

Bike Share Data:

Why Linear Regression is not a good model for this data ?

1. We might get the predicted output (numbers of bikers in any hour) to be –ve
   (Does not make sense!)

Can be addressed by transforming Y into log scale

Regression

$$\log(Y) = \sum_{j=1}^{p} X_j \beta_j + \epsilon.$$

But might loose interpretability in some applications.

2. Assumptions we make in Linear Regression about ϵ:

- zero mean
- constant variance

NOT A GOOD ASSUMPTION
In this case (Why not?)

Generalized Linear Regression

Assumptions we make in Linear Regression about ϵ:

- zero mean
- constant variance

Why this is not a good assumption for bike share problem

Scene 1:

Compute Mean and Variance of numbers of bike users in December, January, February: Between 1 AM - 4:00 AM. (Imagine cold winter weather, midnight hours, and bikers)

Mean: 5.05

Variance: 13.91

Scene 2:

Compute Mean and Variance of numbers of bike users in April, May, June: Between 7 AM - 10:00 AM. (Imagine Spring, Summertime and before noon!)

Mean: 243.59

Variance: 131.7

Conclusion: Linear regression is not a Good assumption

Q. Would Mean be zero and Variance in Scene 1 and Scene 2 be the same?

Generalized Linear Regression

Bike Share Problem: What do we do if not Linear Regression?

Model Y as a Poisson's Distribution

Useful when we are modeling occurrence of events in a given unit of time, distance, space etc. (Discrete events)

Some examples:

Number of average bike users in an hour.

Numbers of car accidents in a day in some city.

Given: Y is a random variable that takes non-negative integers: Y = {0,1,2 …},

And

If: Y follows Poisson's Distribution,

Then:

$$\lambda > 0$$
$$\lambda = E(Y) = Var(Y)$$

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots.$$

Generalized Linear Regression

Bike Share Problem: What do we do if not Linear Regression?

- We re-phrase the problem statement as:
- What is the probability that at any given hour the numbers of bike users is exactly k?

i.e. Pr(Y = k) = ?

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots.$$

So, if we know λ, we can compute the probability.

- Once we can calculate the maximum probability, then we can answer our original question:
- What is the numbers of bike users in any given hour? (Value of k for which probability is maximum)

Note: How we changed seemingly regression problem into classification

Generalized Linear Regression

Bike Share Problem: What do we do if not Linear Regression?

If we know λ, we can compute the probability.

- So, we model λ as a function of or inputs as :

$$\log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

We can think of this as:

- In Linear regression, the relation between target and inputs are linear and variance is constant.
- In this case: Relation between log of Var(target(Y)) and inputs are linear, i.e. log of Variance changes linearly with Inputs

Equivalently,

$$\lambda(X_1, \ldots, X_p) = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}.$$

Generalized Linear Regression

Bike Share Problem: What do we do if not Linear Regression?

If we know λ, we can compute the probability.
If we know β , we can compute λ.

If we know λ, we can compute the probability.
If we know β , we can compute λ.

- Given n training data, We compute β, such that the following likelihood is maximized

$$\ell(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!}$$

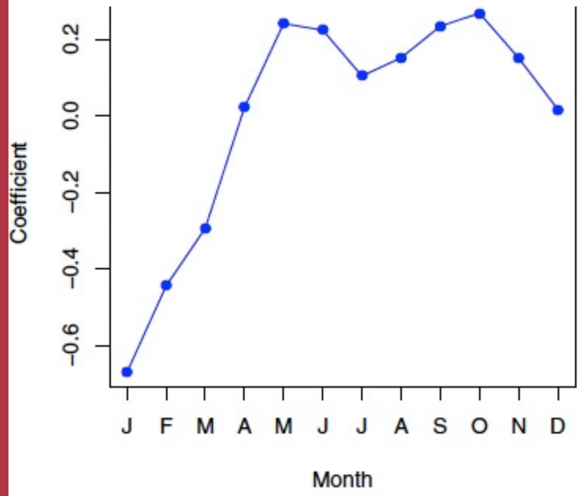
Where, $\lambda(x_i) \; = \; e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}$

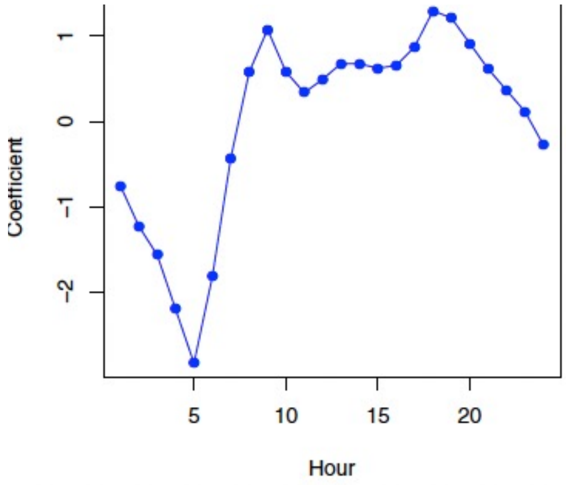
Generalized Linear Regression

Bike Share Data: Poisson's Regression Results:

Table 4.11

	Coefficient	Std. error	z-statistic	p-value
Intercept	4.12	0.01	683.96	0.00
workingday	0.01	0.00	7.5	0.00
temp	0.79	0.01	68.43	0.00
weathersit[cloudy/misty]	-0.08	0.00	-34.53	0.00
weathersit[light rain/snow]	-0.58	0.00	-141.91	0.00
weathersit[heavy rain/snow]	-0.93	0.17	-5.55	0.00





Coeff. Cloudy/Misty = -0.08

Interpretation: (Example)

- Change in weather from clear (baseline) to Cloudy/Misty is associated with change in mean bike users by a **factor** of exp(-0.098) = 0.923.
- In other words: On average, Only 92.3% as many people will use bikes when its cloudy/misty compared to when its clear.

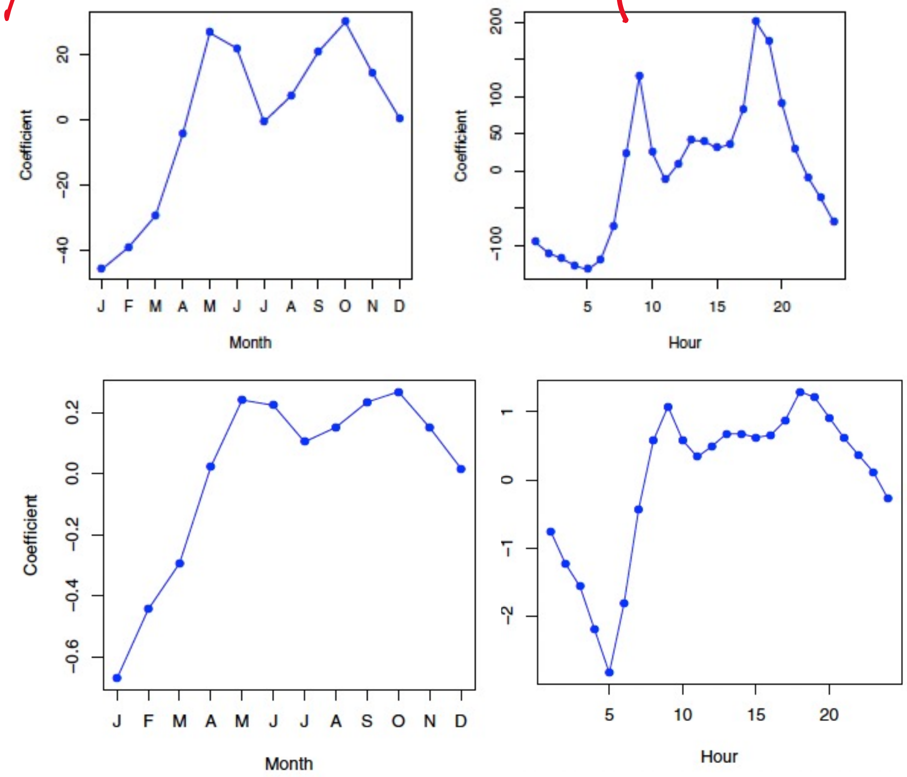
(Try to compare how this interpretation β $(coeff)$ is different in Linear Regression)

Generalized Linear Regression

Comparison: Linear Regression and Poisson's Regression on Bike Share Data

	Coefficient
Intercept	73.60
workingday	1.27
temp	157.21
weathersit[cloudy/misty]	-12.89
weathersit[light rain/snow]	-66.49
weathersit[heavy rain/snow]	-109.75

Coefficient
4.12
0.01
0.79
-0.08
-0.58
-0.93



Top: Linear Regression
Bottom: Poisson's Regression

Note: How the signs of coeffs are consistent and curves looks similar.

Q. Then, what different information we get by doing Poisson's Regression instead of Linear ?

Generalized Linear Regression

Bike Share Data: Poisson's Regression

Different Information

- Interpretation of coefficients are different than in Linear Regression. (Recall our discussion of coefficients)

- Mean-Variance Relationship:
  - Recall: Poisson's Distribution assumes, $\lambda = E(Y) = Var(Y)$
  - i.e. We assumed mean bike users at given hour is equal to variance of users in that hour.
  - Unlike in Linear Regression, where variance is constant and independent of mean.

- Non-negative Fitted (predicted) values.
  (How was this insured?)

Generalized Linear Regression

So,

Q. What is Generalized Linear Regression?

Q. What did we Generalized?

Until now, We discussed:

Linear Regression:
Assumption: Y ~ Gaussian

Logistic Regression:
Assumption: Y ~Binomial

Poisson's Regression:
Assumption: Y ~ Poisson's

Common/General to All:

- All of these Distribution belong to exponential family. (Recall definition of each distribution!)
- We transformed the Predictor (Y) using some transformation (none in Linear Regression, Logits in Logistic Regression and log of Expectation of Y in Poisson's to express relation as linear to input X.)
- We can use similar approach for many other distributions (Gamma, negative binomial, etc) that belongs to exponential family

Hence, All of these methods could be seen as Generalized Linear Regression.