Amir H Gandomi; PhD
Assistant Professor
Stevens Institute of Technology
a.h.gandomi@stevens.edu

Multivariate Data Analysis – BIA 652

Class 6 – Logistic Regression & Naïve Bayes

Overview of Class 6

- Continue Classification – Chapter 11 & 12
  - Logistic Regression
  - Naïve Bayes Classification
  - Ensemble

- Homework:
- Assignment for Oct 18 class: Do A Problem using Logistic Regression, Linear Discriminant Analysis, kNN, and Naïve Bayes Classifications as well as ensembles of these learners.

Logistic Regression

Bayes Rule
Posterior Probability for Discriminant Analysis

Bayes Theorem:

$P(1 \mid X) = q_1 f(X \mid 1) / (q_1 f(X \mid 1) + q_2 f(X \mid 2))$

$= P(X \mid 1)P(1)/P(X)$

Let P_Z be the posterior probability that an observation belongs to population 1. Then:

$P_Z = 1 / (1 + e^{(C - Z)})$

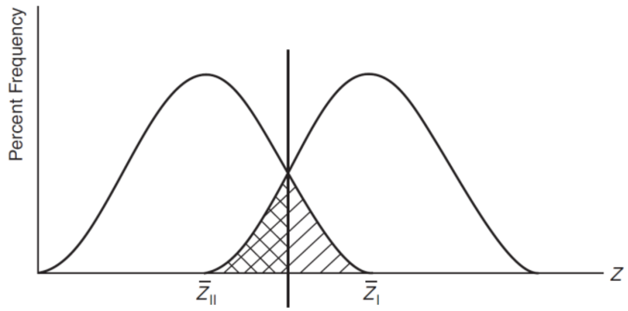
C is a function of prior probabilities

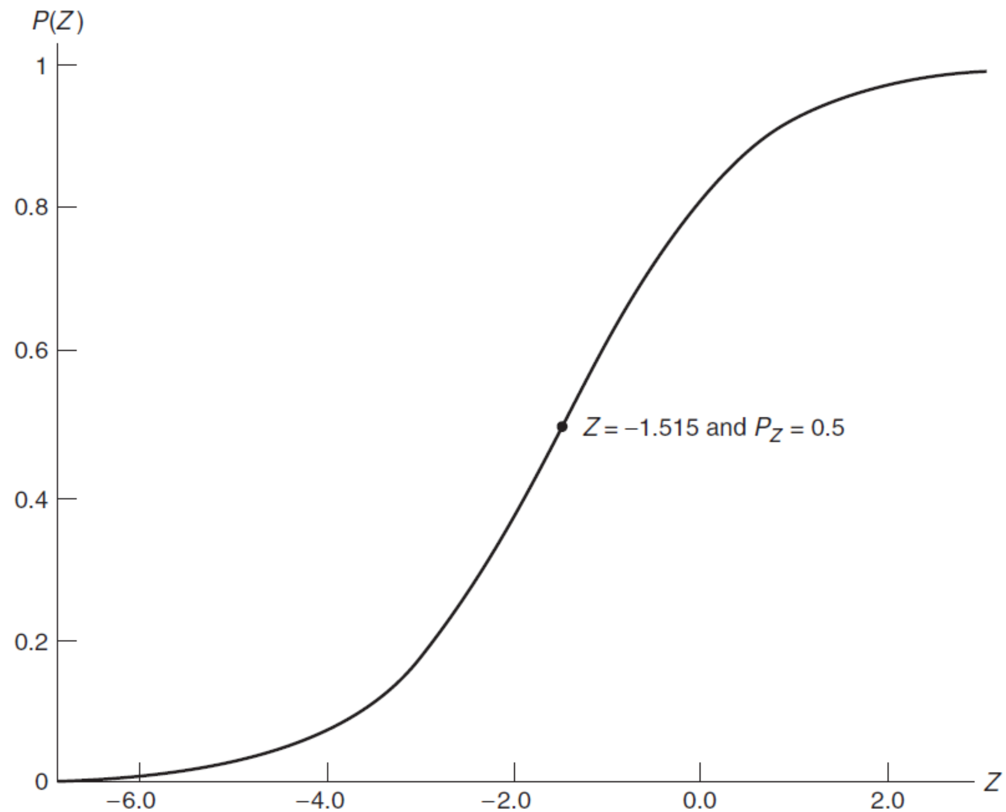
Bayes Rule
Posterior Probability for Discriminant Analysis, Graphically (p 271)

FIGURE 12.1

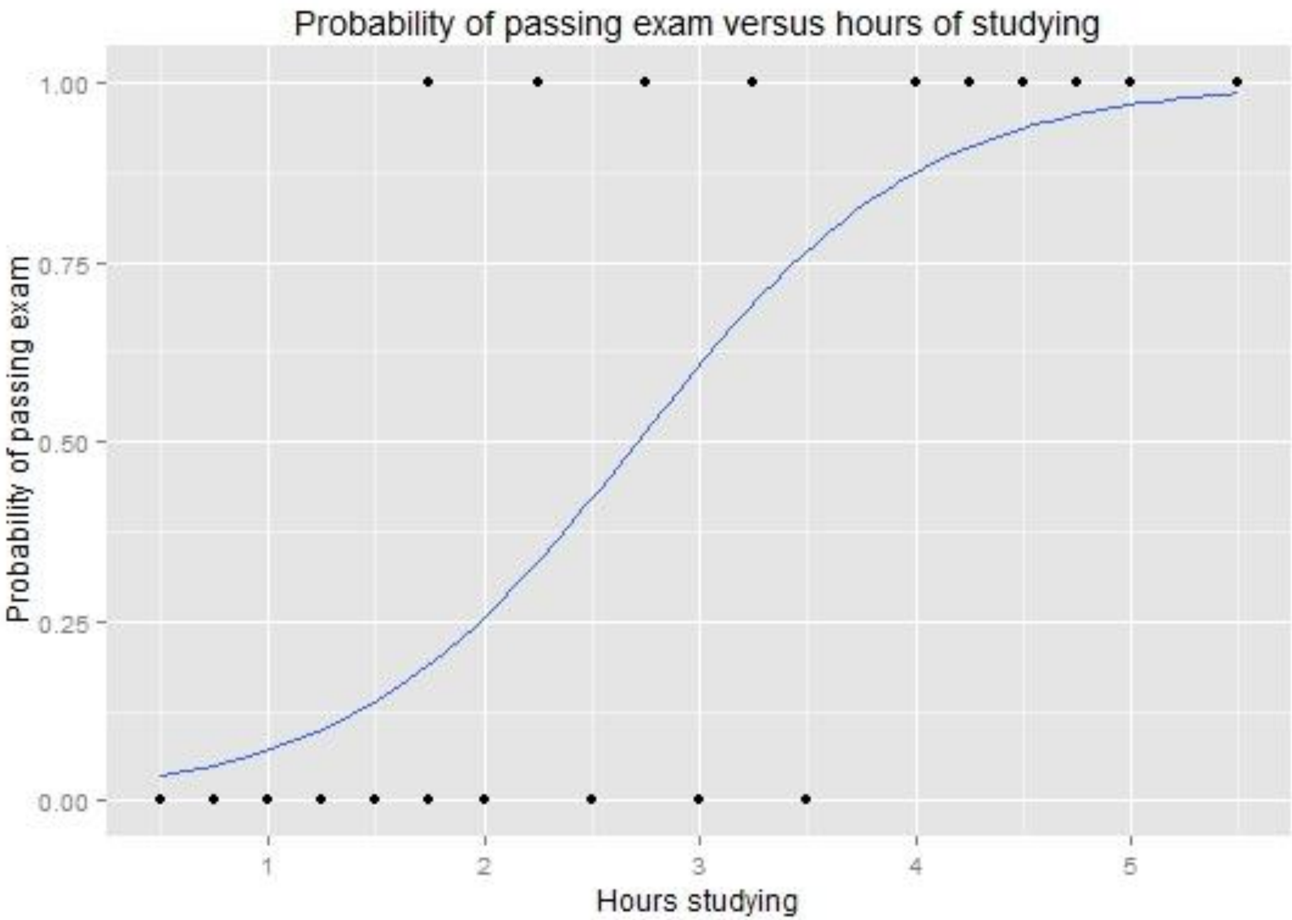
Logistic Function for the Depression Data Set

This is a Logistic Function – Hence Logistic Regression





Logistic Regression



Logistic Regression - Goals

- Similar to Discriminant Analysis
- Classify individuals into one of two groups when:
  - Some of the classifying variables are categorical
- Quantify risk of outcome
- Test which variables are useful
- http://www.ats.ucla.edu/stat/sas/seminars/sas_logistic/logistic1.htm

Logistic Regression – Model & A little basic probability

- Start with posterior probability from discriminant analysis
- The probability of belonging to Group 1 is a logistic function:

$$P_z = \frac{1}{1+e^{C-Z}}$$

- If P is the probability of an event
- The odds of that event are:

$$Odds = \frac{P}{(1 - P)}$$

Risk Ratio and Odds Ratio

- The ratio of these two probabilities R1/R2 is the relative risk or risk ratio (RR):

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}}$$

- If O1 is the odds of event in the Treatment group and O2 is the odds of event in the control group then the odds ratio (OR) is O1/O2:

$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}}$$

- Just like the RR, OR is a way of measuring the effect of the tutoring program on the odds of an event.

Ref.: https://www.theanalysisfactor.com/the-difference-between-relative-risk-and-odds-ratios/

Logistic Regression
Example: depression vs. gender (p 273)

Table 12.1: Classification of individuals by depression level and sex

	Depression		
Sex	Yes	No	Total
Female (1)	40 **(a)**	143 **(b)**	183
Male (0)	10 **(c)**	101 **(d)**	111
Total	50	244	294

$$RR = \frac{\text{Risk of event in the Treatment group}}{\text{Risk of event in the Control group}} = \frac{a/(a+b)}{c/(c+d)}$$

$$OR = \frac{\text{Odds of event in Treatment group}}{\text{Odds of event in Control group}} = \frac{a/b}{c/d}$$

Logistic Regression
Risk Ratio

	D	ND	Total
E(Female)	a (40)	b (143)	a + b (183)
NE(Male)	c (10)	d (101)	c + d (111)
Total	a + c (50)	b + d (244)	(294)

RR = P(D|E) /P(D|NE) = (a/(a+b)) / (c / (c+d))

e.g. RR = P(D|E) /P(D|NE) = (40/183)/(10/111) = 0.219/0.090 = 2.43

Logistic Regression
Odds Ratio

	D	ND	Total
E(Female)	a (40)	b (143)	a + b (183)
NE(Male)	c (10)	d (101)	c + d (111)
Total	a + c (50)	b + d (244)	(294)

OR = [odds (D|E)] / [odds (D|NE)]

OR =[P(D|E) /P(ND|E)] / [P(D|NE) /P(ND|NE)] =

[(a/(a+b)) / (b / (a+b))] / [(c/(c+d)) / (d/ (c+d))] = ad / bc

e.g. OR = (0.219/0.781) / (0.09/ 0.91) = (40 * 101) / (10 * 143) = 2.83

Logistic Regression - Model

- P_Z = Logistic Function = P(D|X) = P(1|X)

$$P_Z = \frac{1}{1+e^{C-Z}}$$

- It can be shown that:

$$\text{Odds} = \frac{P_Z}{(1-P_Z)} = e^{(Z-C)}$$

- Therefore: ln (odds) = Z – C
- And: Z is a linear function of X.

Logistic Regression – Model
Odds are multiplicative

- ln (odds) = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
- Odds = $\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)$

  $= (e^{\beta_0})\ (e^{\beta_1 X_1})(\ldots)(e^{\beta_p X_p})$

- Or: Odds = $\text{constant}_0 \times \exp(\text{constant}_1 \times X_1)$

  $\vdots$

  $\times \exp(\text{constant}_p \times X_p)$

Logistic Regression – Model
Log (Odds) are additive

- $\ln(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = \text{Z-C}$

- Or: $\ln(\text{Odds}) = \text{constant}_0 + (\text{constant}_1 \times X_1)$

$$\vdots$$

$$+ (\text{constant}_p \times X_p)$$

Logistic Regression – Model Log (Odds) are additive

$Ln\ (Odds)$ is called logit

$$Ln\ (Odds) = Ln\left(\frac{P_z}{1 - P_z}\right) = \boldsymbol{\beta X} = Z - C$$

β is coefficient vector and X is variable vector

It can be presented by probability (P), as:

$$P_z = \frac{1}{1 + e^{C-Z}} = \frac{1}{1 + e^{-\boldsymbol{\beta X}}}$$

Logistic Regression – Model Flexibility

- X (independent) variables can be continuous or categorical
- Interactions can be incorporated
- Coefficients are estimated by maximum likelihood
- Most computer programs implicitly use prior probabilities estimated from sample.

Logistic Regression – Model Example: circulatory shock

- Patients are in shock
- Outcome = survival or death
- Use Discriminant Function Analysis or Logistic Regression to identify risk factors for death

Logistic Regression – Model Example: circulatory shock Parameter Estimates

- Discriminant Function can be used to estimate parameters
- Better to use Maximum likelihood estimates
- Depends on the idea of a Generalized Linear Model (GLM)

Logistic Regression – Model Generalized Linear Model (GLM)

- Logistic regression is an example of the GLM
- Define Y = outcome = 1 (event) or 0 (not)
- E (Y | X's) = μ = P(1 | X's)
- Find a function g(μ), called the *link* function, such that:

  g(μ) = linear function of the X's
- This is called the GLM
- Here we take g(μ) = ln (odds) = logit function

Logistic Regression – Model Estimation

- Model is: $g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
- Need to estimate: $\beta_0, \beta_1, \beta_2, \ldots \beta_p$
- Use an interactive process called Iterative Weighted Least Squares:
  1. Start with initial estimates of parameters
  2. Evaluate the score equations (derivative of log-likelihood = 0)
  3. Solve the score equations and get new estimates of parameters
  4. Repeat until convergence.

Logistic Regression – Model Example: Depression Data Set

- Model is: $g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
- Need to estimate: $\beta_0, \beta_1, \beta_2, \ldots \beta_p$
- X's = Age, Income, Sex (0 if Male and 1 if Female)
- Results:
  - $-C = \beta_0 = -0.676$
  - $Z = \beta_1 \text{ Age} + \beta_2 \text{ Income} + \beta_3 \text{ Sex}$
    $= -0.021 \text{ Age} - 0.037 \text{ Income} + 0.929 \text{ Sex}$
- Interpretation:
  - ❖ The probability of being depressed decreases with age and income, but increases if female

Logistic Regression – Model Example: Depression Data Set Interpretation – Binary Variable

Y = outcome is binary: 0, 1 variable (e.g. depressed = 1)

X is binary: 0, 1 variable

- Odds (Y = 1 | X) = $e^{\alpha + \beta X + \text{other X's}}$
- Odds (Y=1 | X=1) = $e^{\alpha + \beta + \text{other X's}}$
- Odds (Y=1 | X=0) = $e^{\alpha + \text{other X's}}$
- OR = Odds (Y=1 | X=1)/ Odds (Y=1 | X=0) = e^{β}
- Ln(OR) = β

Logistic Regression – Model Example: Depression Data Set Odds Ratio – Binary Example

- Sex – Categorical:
  - $e^{(0.929)} = 2.582$ = adjusted odds ratio if female
  - Unadjusted OR (Computed before)
  - OR = [odds (D | E)] / [odds (D | NE)] = 2.83

Logistic Regression – Model Example: Depression Data Set Interpretation – Continuous Variable

- Y = outcome is binary: 0, 1 variable (e.g. depressed = 1)
- X is continuous: e.g. age

- Compare X to X+1 (1 year older)
- Odds (Y = 1 | X+1) = $e^{\alpha + \beta X + \beta + \text{other } X\text{'s}}$
- Odds (Y=1 | X) = $e^{\alpha + \beta X + \text{other } X\text{'s}}$
- OR = Odds (Y=1 | X+1)/ Odds (Y=1 | X) = e^{β}
- Ln(OR) = β

Logistic Regression – Model Example: Depression Data Set Odds Ratio – Continuous Example

- Age – Continuous:
  - $e^{(-0.021)} = 0.98$ = adjusted incremental odds ratio for increase of 1 year of age
  - What about a 10 year increase?
    - $e^{(10 \times -0.021)} = 0.81$

Logistic Regression – Model Example: Depression Data Set Odds Ratio – Nominal Variable > 2 Categories

- Y = outcome is binary: 0, 1 (e.g. depressed = 1)
- X nominal: (e.g. religion)
- Coefficient of D_1 = ln (OR) for "Catholic" vs. "Other"

Religion	D_1	D_2	D_3
Catholic	1	0	0
Protestant	0	1	0
Jewish	0	0	1
Other	0	0	0

Logistic Regression – Model Example: Depression Data Set Adjusted Risk Ratio

- $RR = P(Y=1 \mid X=1)/P(Y=1 \mid X=0)$
- $P(Y=1 \mid X) = e^{LC} / (1 + e^{LC})$
  - Where $LC = A + B_1 X_1 + \ldots$

- Example: Depression, X = sex = 1 if F, age = 30, income = 10 ($10K/year)
- Find adjusted RR for F vs. M
- LC = -0.676 – 0.021 Age – 0.037 Income + 0.929 Sex

Logistic Regression – Model Example: Depression Data Set Adjusted Risk Ratio

- LC = -0.676 – 0.021 Age – 0.037 Income + 0.929 Sex
- (LC | X=1) = -0.676 – 0.021(30) – 0.037(10) +0.929(1)
  = - 0.747
- (LC | X=0) = -0.676 – 0.021(30) – 0.037(10) +0.929(0)
  = - 1.676
- P(Y=1 | X=1) = P(Depr | F) = $e^{-0.747}$ / (1 + $e^{-0.747}$) = 0.3215
- P(Y=1 | X=0) = P(Depr | M) = $e^{-1.676}$ / (1 + $e^{-1.676}$) = 0.1576
- Adjusted RR = 0.3215 / 0.1576 = 2.04
- Recall: Unadjusted RR = 2.43, adjusted OR = 2.584, and unadjusted OR = 2.83

Logistic Regression
Types of Observational Studies

	Exposed	Unexposed	Total
Case			
Non-case			
Total			

Case – control Study:
Fixed Margins

Cohort Study:
assumed fixed margins

Cross – sectional study:
fixed grand total

Logistic Regression
Validity of OR and RR

- Type of Observational research methods:
  - Cohort Study
  - Cross-sectional Study
  - Case Control Study

- OR is valid if there is a cohort, cross-sectional or case control study
- RR is valid only if there is a cohort or cross-sectional study

More reading: http://dx.doi.org/10.1136/emj.20.1.54

Logistic Regression
Confounders and Effect Modifiers

- **Confounding**: A situation in which the effect or association between an exposure and outcome is distorted by the presence of another variable.
  - X_2 is a confounder for the effect of X_1 if X_2 is correlated with both Y and X_1 and prediction is distorted by presence of X_2
- **Effect modification**: occurs when the effect of a factor is different for different groups.
  - X_2 is an effect modifier for the effect of X_1 when the effect of X_1 is different for different X_2

- Example:
  - Y = CHD (yes/no),
  - X_1 = risk factor = Diabetes (yes/no),
  - X_2 = Age (old/young)

Logistic Regression Example: Confounders

- OR (CHD vs. Diabetes) = 2
- OR (CHD vs. Diabetes | Old) = 1.5
- OR (CHD vs. Diabetes | Young) = 1.5
- Hence: Include both X_1 and X_2

- **Confounders Check**: In a model with X_1 and X_2 :
  - If coefficient of X_2 is significant, then X_2 is a confounder for X_1
  - If not, use X_1 only

Logistic Regression Example: Effect Modifier

- OR (CHD vs. Diabetes) = 2
- OR (CHD vs. Diabetes | Old) = 3
- OR (CHD vs. Diabetes | Young) = 1.4
- Hence: Include both X_1 and X_2 and $X_1 \times X_2$

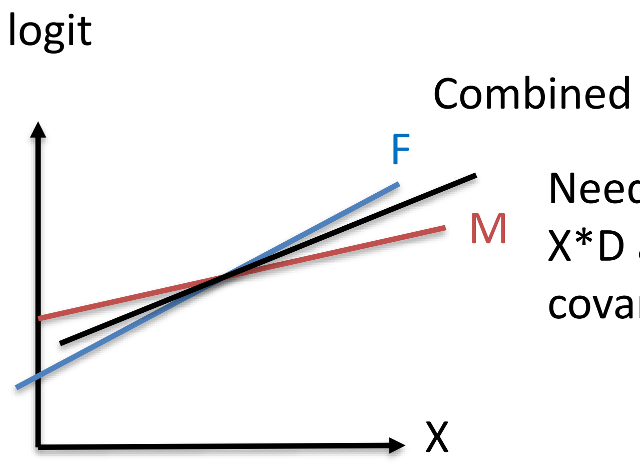
- **Effect Modifier Check**: In a model with X_1 and X_2 and $X_1 \times X_2$
  - If interaction is significant, then X_2 is an effect modifier for X_1 (and vice versa)
  - If interaction is not significant, check for confounding.

Logistic Regression Example: Review

- Y = outcome (binary)
- X = continuous covariate
- D = binary covariate (female/male)
- Logistic Regression = regression of logit on X, D, X*D.

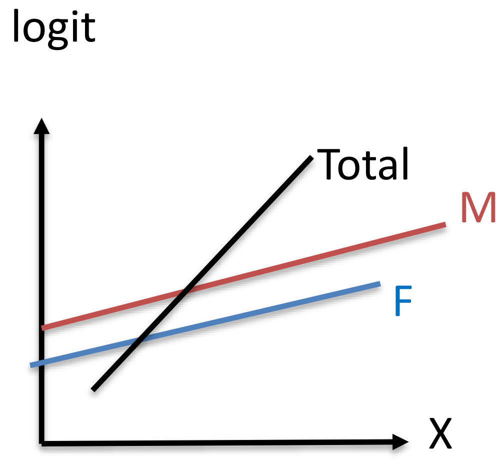
Logistic Regression - Review

Effect Modification



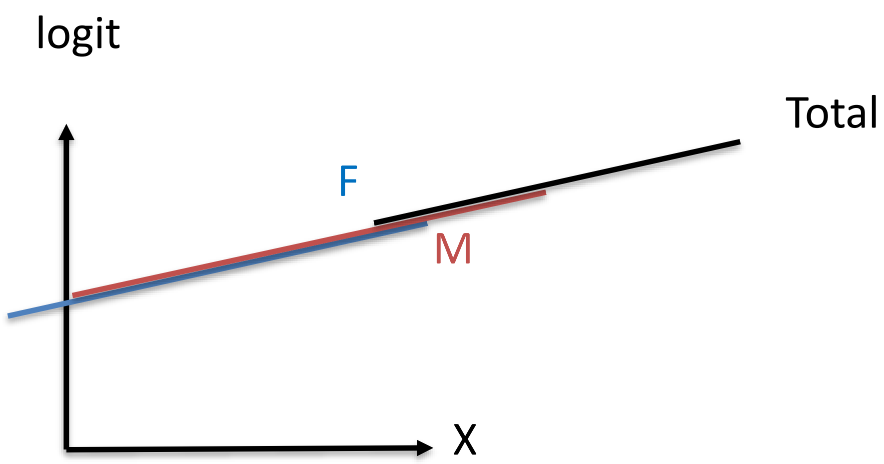
Need X,D and X*D as covariates

Confounding



Need X and D as covariates

Neither Effect Modification nor Confounding



Need X as the only covariate

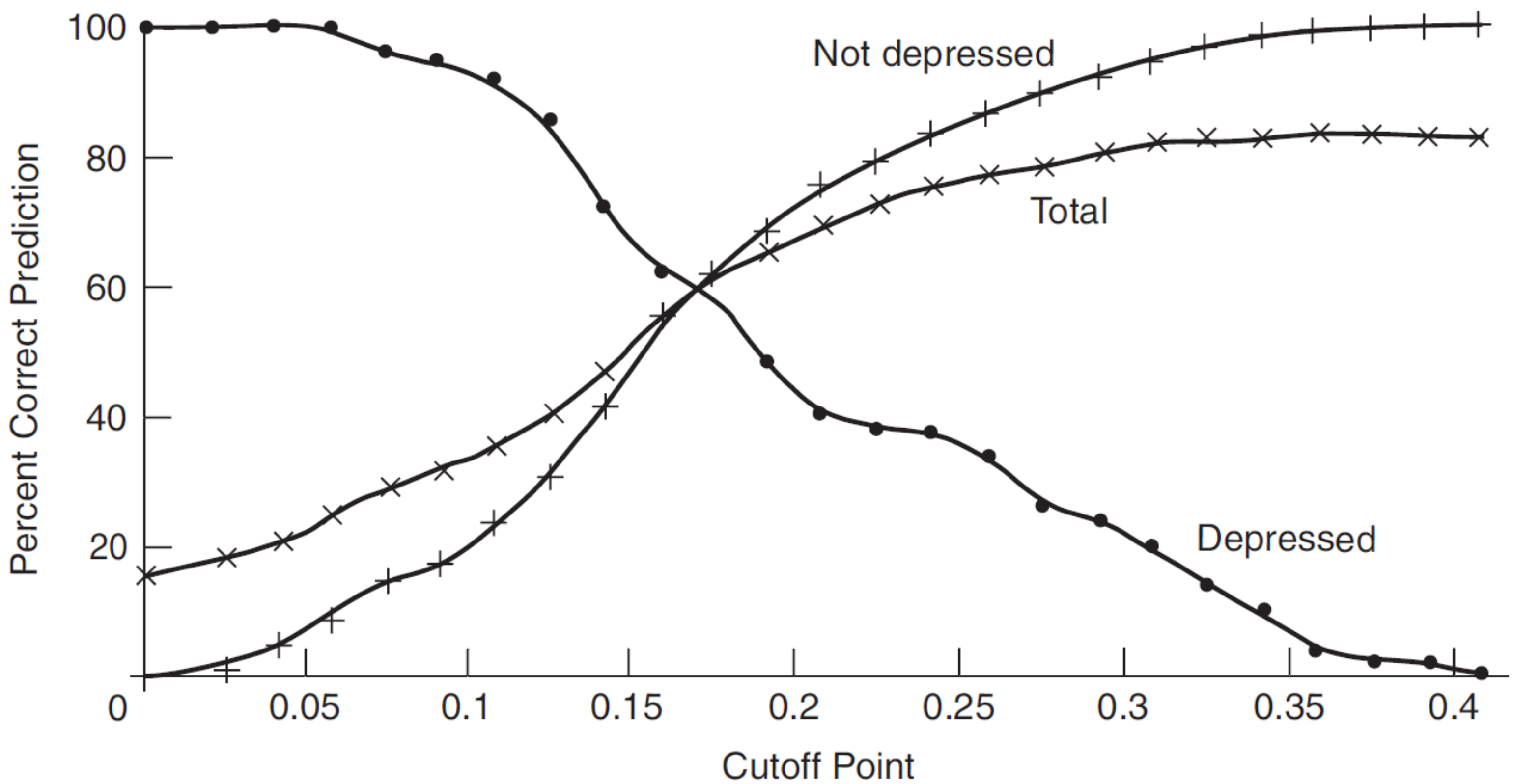
Model Evaluation

- For Classification:
  - Choose cutoff P_C
  - Place individual in group 1 if: $P_Z > P_C$
  - What percent correctly classified?
  - Repeat for many values of P_C

Probability of Correct Classification

FIGURE 12.5

Percentage of Individuals Correctly Classified by Logistic Regression



Nominal (Multinominal) Logistic Regression
(More than 2 Categories)

- Outcome has more than 2 categories
  - Disease: Type 1 Diabetes, Type 2 Diabetes, No Diabetes
  - Disease: Cure, Remission, Sick
  - Security: Free, Alert, Alarm
  - Many: High, Medium, Low

- Choose a reference group and define;
  "odds" = P(cat. #i) / P(ref. category)

Model

- Ln [P(cat. #i) / P(ref. category)]
  $= a_i + \beta_{i1} X_1 + \beta_{i2} X_2 + . \ . \ . + \beta_{ip} X_p$
- The number of parameters with k categories is: (k-1)x(P + 1), as opposed to P + 1 in binary LR
- "odds ratio" of category #i vs reference category = e^{β}

Example (p. 300)

Define: CESD 0-9 = No Depression;
CESD 10-15 = Borderline Depression;
CESD > 15 = Clinical Depression

Table 12.4: Estimated coefficients from nominal logistic regression

Term	Coefficient	Standard error	P-value
Borderline depressed			
Sex (1=Female)	−0.017	0.332	0.96
Age (years)	−0.016	0.009	0.08
Income ($1,000)	−0.017	0.012	0.14
Constant	−0.296	0.755	0.70
Clinically depressed			
Sex (1=Female)	0.925	0.393	0.02
Age (years)	−0.024	0.009	0.01
Income ($1,000)	−0.040	0.014	0.01
Constant	−1.136	0.867	0.19

OR's

- OR (borderline depression vs no depression, F vs M)
  exp (-0.017) = 0.98

- OR (clinical depression vs no depression, F vs M)
  exp(0.925) = 2.52

Caveats

- Training sample must be correctly classified and representative of the population
- Fundamental assumption of logistic model is that the ln(odds) is linearly related to the independent variables
- The coefficient of any one variable can vary widely, depending on what others are included in the model

Naïve Bayes

Classification Naïve Bayes

- Bayesian
  - Traditional Naive Byes: Simple & Naïve
    - A simple probabilistic classifier based on applying Bayes' theorem with strong (naive) **independence assumptions**

$$p(C|F_1, \cdots, F_n)$$

$$\overline{} \rightarrow p(C, F_1, \cdots F_n)$$

$$p(C|F_1, \cdots, F_n) = \frac{p(C)p(F_1, \cdots, F_n|C)}{p(F_1, \cdots, F_n)}$$

constant

$$posterior = \frac{prior \times likelihood}{evidence}$$

Naïve Bayesian Classification

$$p(C|F_1, \cdots, F_n) = \frac{p(C)p(F_1, \cdots, F_n|C)}{p(F_1, \cdots, F_n)}$$

$$\begin{aligned} p(C, F_1, \cdots, F_n) &= p(C)p(F_1, \cdots, F_n|C) \\ &= p(C)p(F_1|C)p(F_2, \cdots, F_n|C, F_1) \end{aligned}$$

Using Induction:

$$= p(C)p(F_1|C)p(F_2, |C, F_1)p(F_3, \cdots, F_n|C, F_1, F_2)$$

$$\cdots p(F_n|C, F_1, F_2, \cdots F_n)$$

Naïve Bayesian Classification

$$p(C, F_1 \cdots, F_n) = p(C)p(F_1, \cdots, F_n|C)$$
$$= p(C)p(F_1|C)p(F_2, \cdots, F_n|C, F_1)$$

Using Induction:

$$= p(C)p(F_1|C)p(F_2, |C, F_1)p(F_3, \cdots F_n|C, F_1, F_2) \cdots p(F_n|C, F_1, F_2, \cdots F_{n-1})$$

Naïve Assumption: $p\left(F_i \middle| C, F_j\right) = p(F_i|C)$

$$p(C, F_1, \cdots, F_n) = p(C)p(F_1|C)\, p(F_2, |C) \cdots$$

or

$$= p(C) \prod_1^n p(F_i|C)$$

Naïve Bayesian Classification

$$p(C, F_1 \cdots F_n) = p(C)p(F_1|C)\, p(F_2, |C) \cdots$$

or

$$= p(C) \prod_1^n p(F_i|C)$$

$p(C)$ **= % of Class C in Training Set**

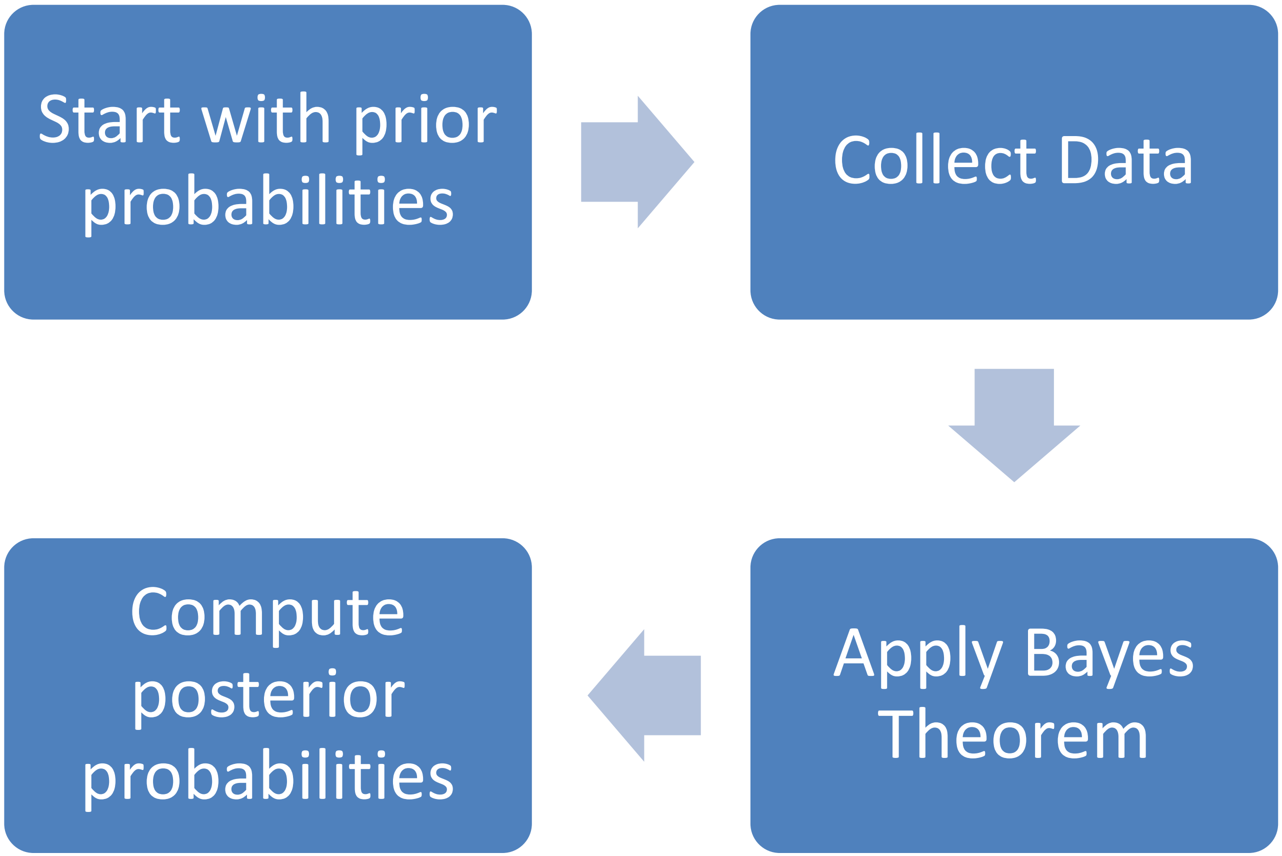
$p(F_i|C)$ **= Probability of F_i Given C in Training Set**

Goal: Classify a New Occurrence, using Max a-posteriori

That is:

Classify $(f_1, \cdots, f_n) = argmax\, p(C = c) \prod_1^n p(F_i = f_i|C = c)$

Naïve Bayes Approach



Naïve Bayesian Classification

Example(Sex Classification)

Training

Example training set below.

sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Parameters estimation:

Probability distribution of every feature in every class

The class priors:

P(male)= P(female) = 0.5.

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Naïve Bayesian Classification

Example(Sex Classification)

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

$$\begin{aligned}\text{evidence} = {} & P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male}) \\ & + P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})\end{aligned}$$

Naïve Bayesian Classification

Example(Sex Classification)

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

P(male) = 0.5

P(height | male) = 1.5789 (A probability density greater than 1 is OK. It is the area under the bell curve that is equal to 1.)

P(weight | male) = 5.9881e-06

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789$$

P(foot size | male) = 1.3112e-3

posterior numerator (male) = their product = 6.1984e-09

Naïve Bayesian Classification

Example(Sex Classification)

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

P(female) = 0.5

P(height | female) = 2.2346e-1

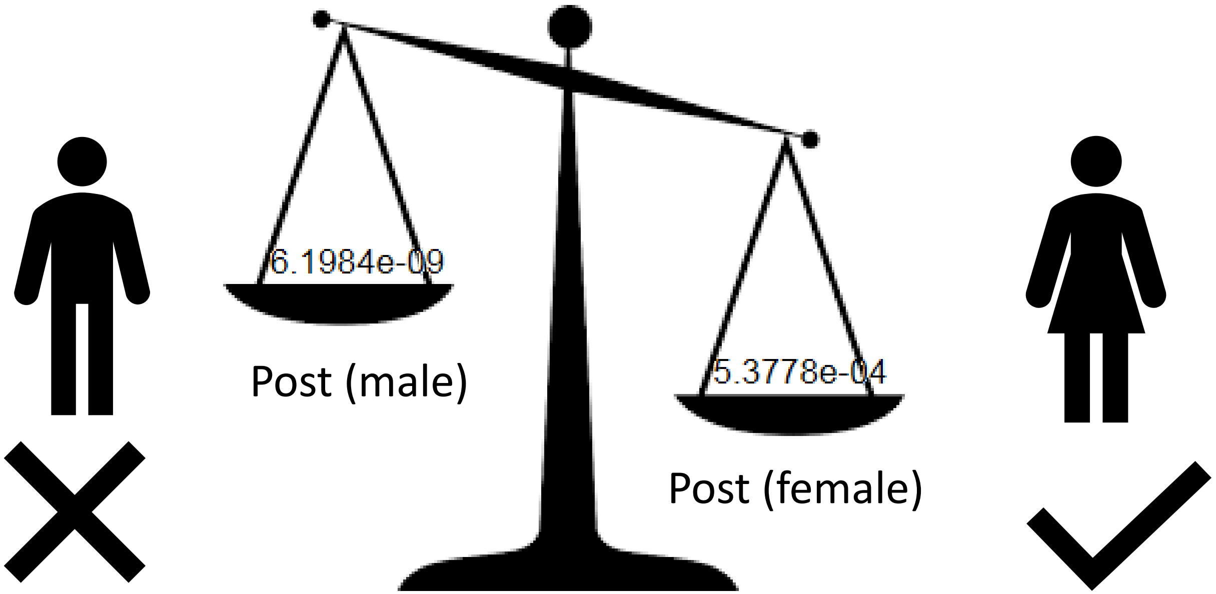
P(weight | female) = 1.6789e-2

P(foot size | female) = 2.8669e-1

posterior numerator (female) = their product = 5.3778e-04

Naïve Bayesian Classification

Example (Sex Classification)



Ensembles

Classification of Classifiers

Supervised Classifiers

Discriminative

Learn differences

Model P(y|x)

Such as: LR, KNN

Generative

Learn each class

Model P(x,y)

Such as: NB

Discriminative classifiers usually perform better when we have enough data

Ensemble methods - 1

- **Simple Ensemble methods**
  - **Committees:**
    - Majority Vote

  - **Weighted Average:**

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \dots, x_m) \\ y_2 &= f_2(x_1, x_2, \dots, x_m) \\ &\dots \\ y_n &= f_2(x_1, x_2, \dots, x_m) \end{aligned} \quad \rightarrow \quad y_e = \sum_{i=1}^{n} \omega_i y_i$$

    up weight better predictors

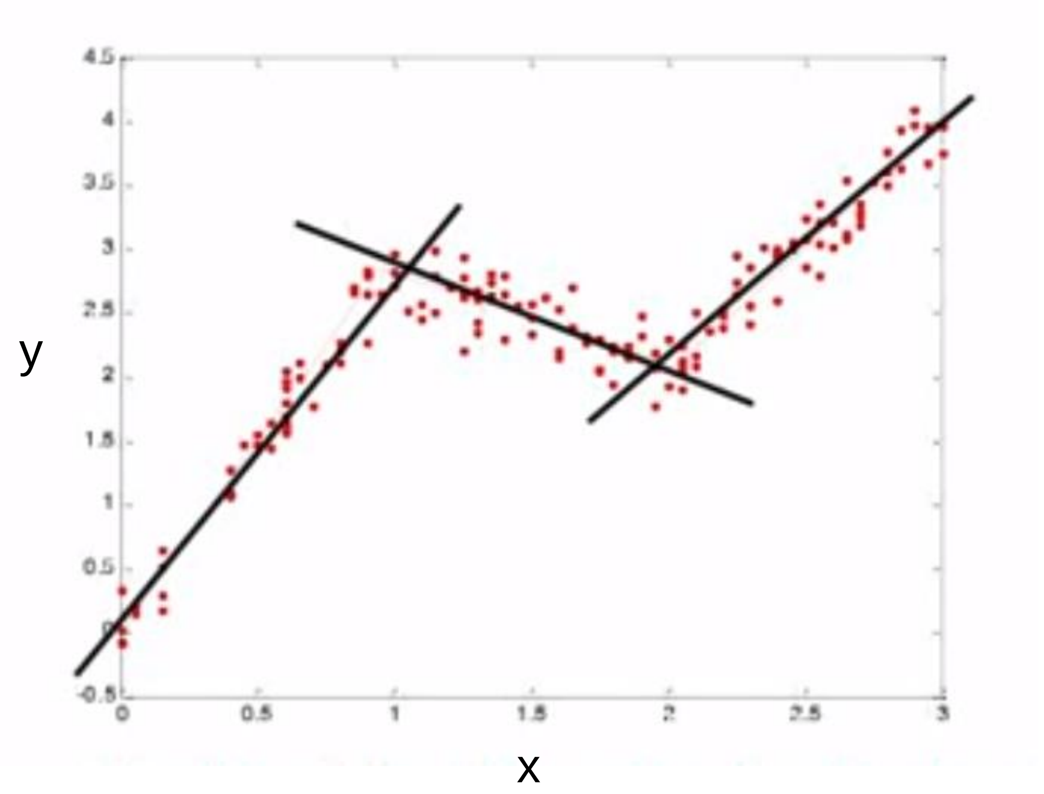
  - **One Option:**

$$\begin{aligned} y_1 &= f_1(x_1, x_2, \dots, x_m) \\ y_2 &= f_2(x_1, x_2, \dots, x_m) \\ &\dots \\ y_n &= f_2(x_1, x_2, \dots, x_m) \end{aligned} \quad \rightarrow \quad y_e = f_e(y_1, y_2, \dots, y_n)$$

    if f_e is linear it is similar to weighted average

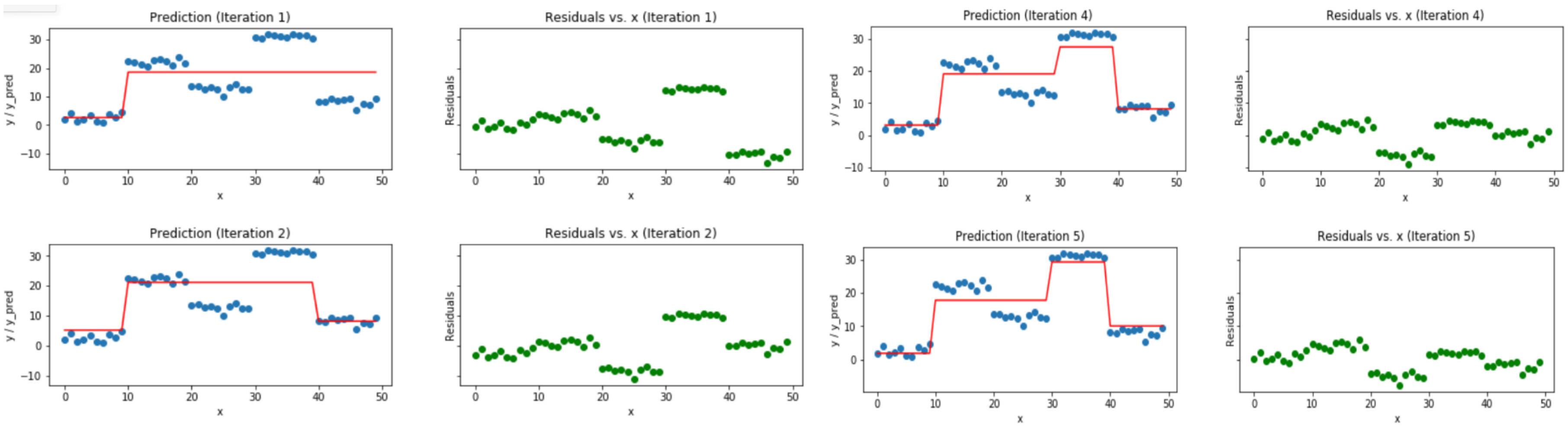
Ensemble methods - 2

- **Mixture of Expert**
  - **Example: mixture of three linear predictor experts**

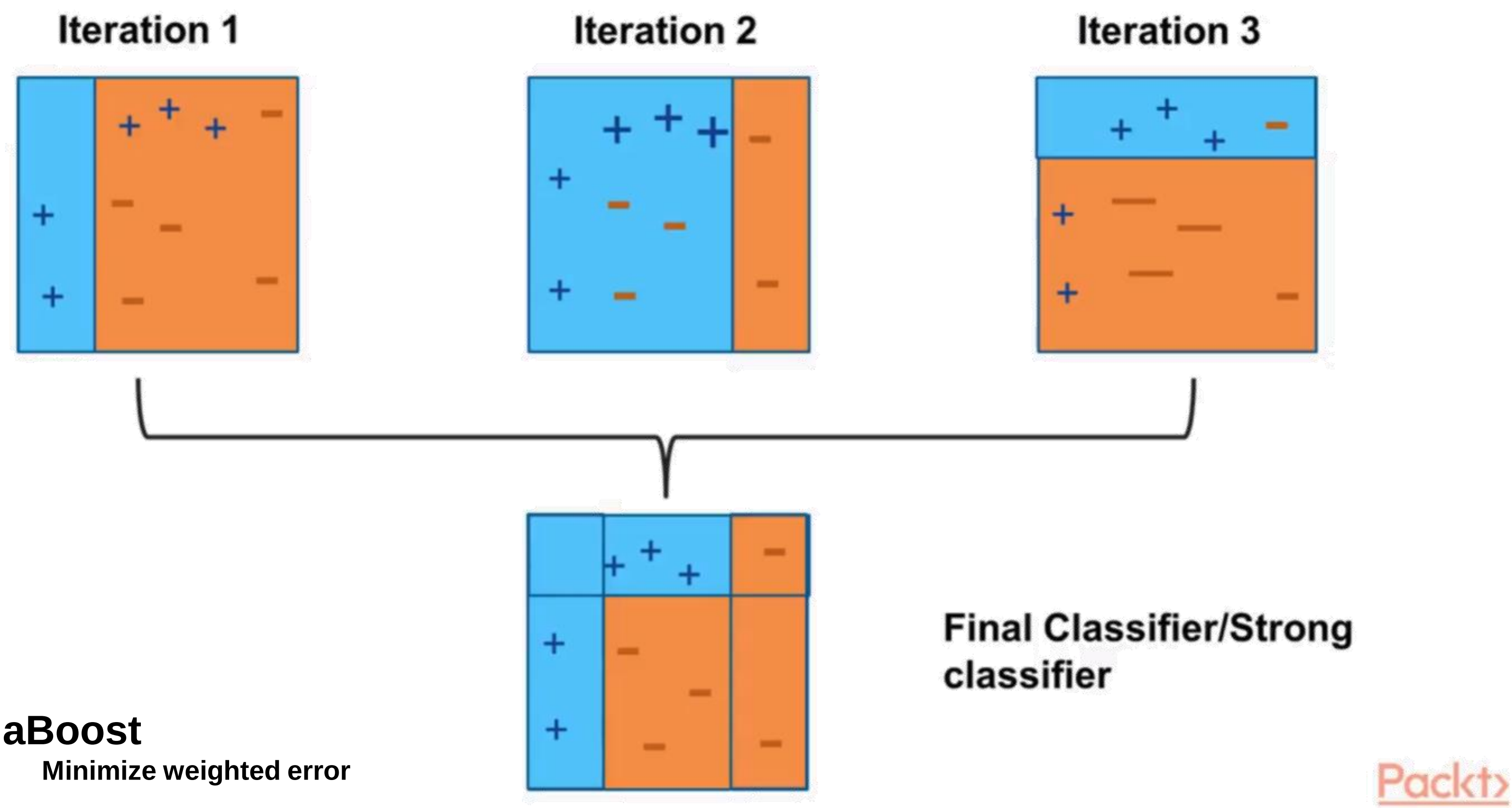


Ensemble methods - 3

- Other methods
  - Bagging
    - Learn many classifier, each with only part of the data
    - Combining models (e.g. averaging)
  - Gradient Boosting
    - Learn to predict the residual
    - XGBoost (A successful gradient boosting)
    - Combining giving a better predictor, Can try to correct its errors also,& repeat



Ensemble methods - 4



AdaBoost

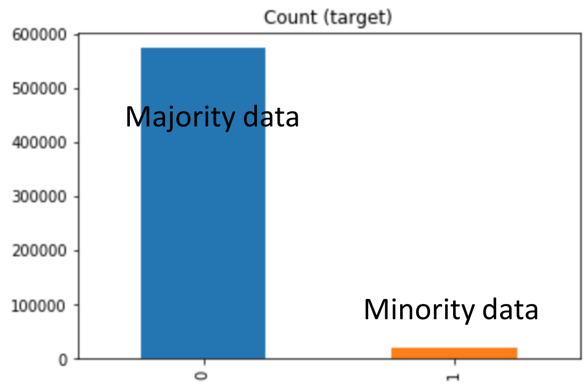
Minimize weighted error

Imbalanced Data Issue

- **Introduction**

Imbalanced data issue usually refers to classification problems when we have unequal instances for different classes. We'll have large mount of data for one class (majority data) and much fewer data for one or more other class(minority data)

```
Class 0: 573518
Class 1: 21694
Proportion: 26.44 : 1
```



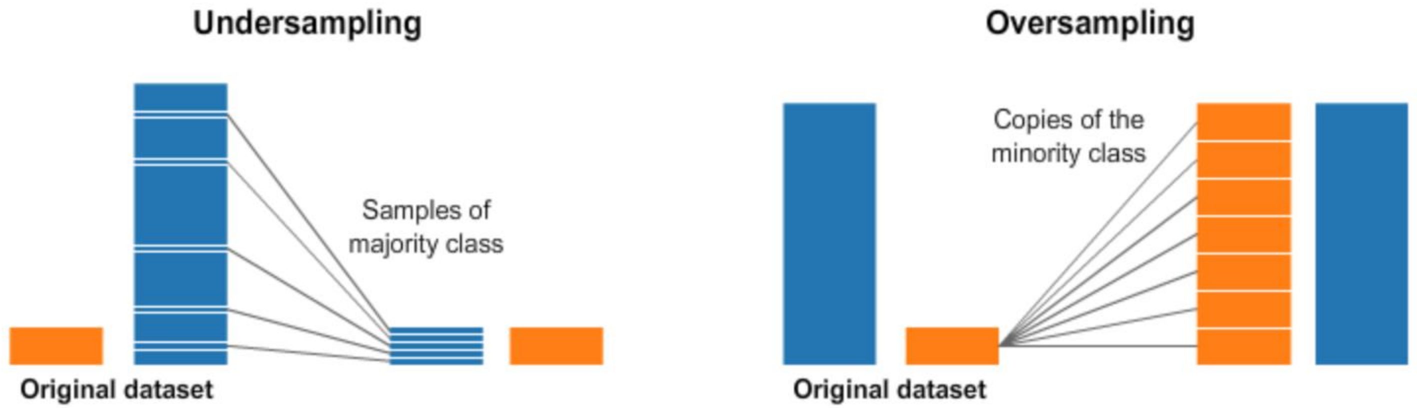
- **Challenge**

The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets. For example logistic regression tend to have bias towards classes which have number of instances. It will only predict majority data and treat minority data as noise.

How to handle the Imbalanced data

- **Oversampling and Undersampling**

A widely adopted technique for dealing with highly unbalanced datasets is called resampling. It consists of removing samples from the majority class (under-sampling) and / or adding more examples from the minority class (over-sampling).



How to handle the Imbalanced data

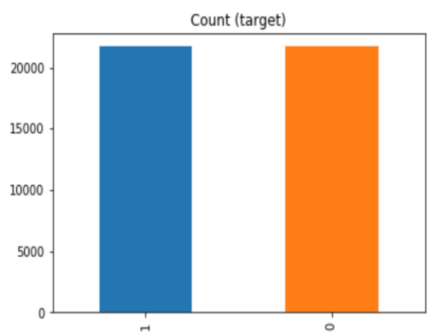
- Python code for oversampling and undersampling

Undersampling

```
# Class count
count_class_0, count_class_1 = df_train.target.value_counts()

# Divide by class
df_class_0 = df_train[df_train['target'] == 0]
df_class_1 = df_train[df_train['target'] == 1]
```

```
df_class_0_under = df_class_0.sample(count_class_1)
df_test_under = pd.concat([df_class_0_under, df_class_1], axis=0)

print('Random under-sampling:')
print(df_test_under.target.value_counts())

df_test_under.target.value_counts().plot(kind='bar', title='Count (target)');
```

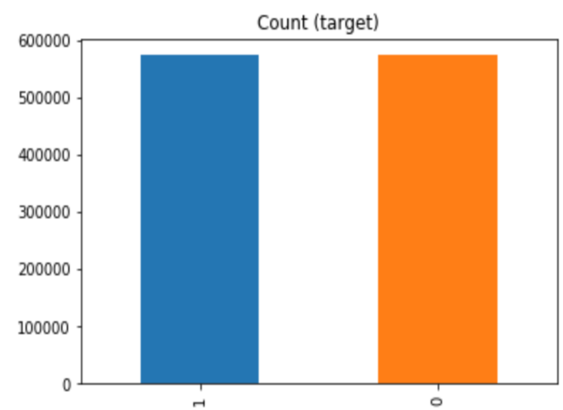
```
Random under-sampling:
1    21694
0    21694
Name: target, dtype: int64
```



Oversampling

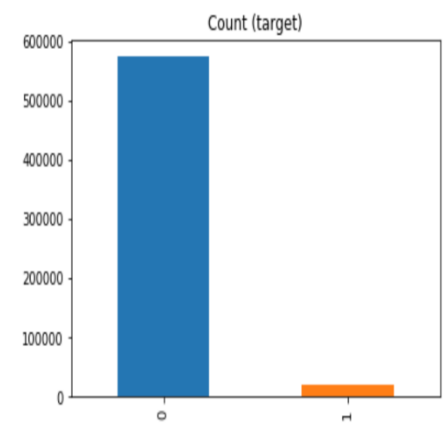
```
df_class_1_over = df_class_1.sample(count_class_0, replace=True)
df_test_over = pd.concat([df_class_0, df_class_1_over], axis=0)

print('Random over-sampling:')
print(df_test_over.target.value_counts())

df_test_over.target.value_counts().plot(kind='bar', title='Count (target)');
```

```
Random over-sampling:
1    573518
0    573518
Name: target, dtype: int64
```



Original data size

```
Class 0: 573518
Class 1: 21694
Proportion: 26.44 : 1
```



STEVENS
INSTITUTE of TECHNOLOGY
School of Business

stevens.edu

Amir H Gandomi; PhD
Assistant Professor of Information Systems
a.h.gandomi@stevens.edu