



Inspire...Educate...Transform.

Statistics and Probability in Decision Modeling

Logistic Regression, ROC and AUC and Naïve Bayes

Dr. Venkatesh Sunkad

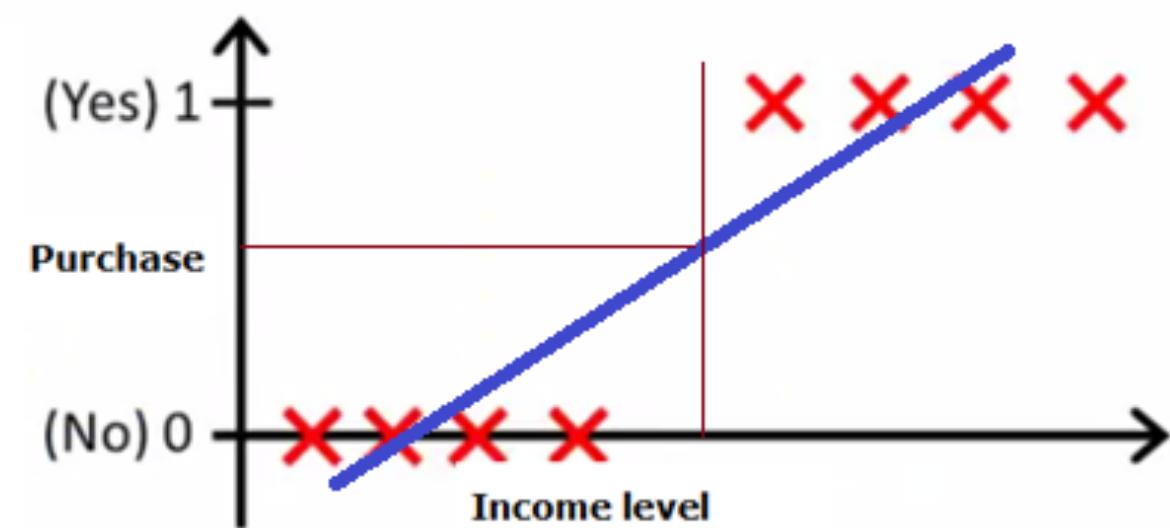
Jan 18 2020

slides courtesy Dr. Sridhar Pappu

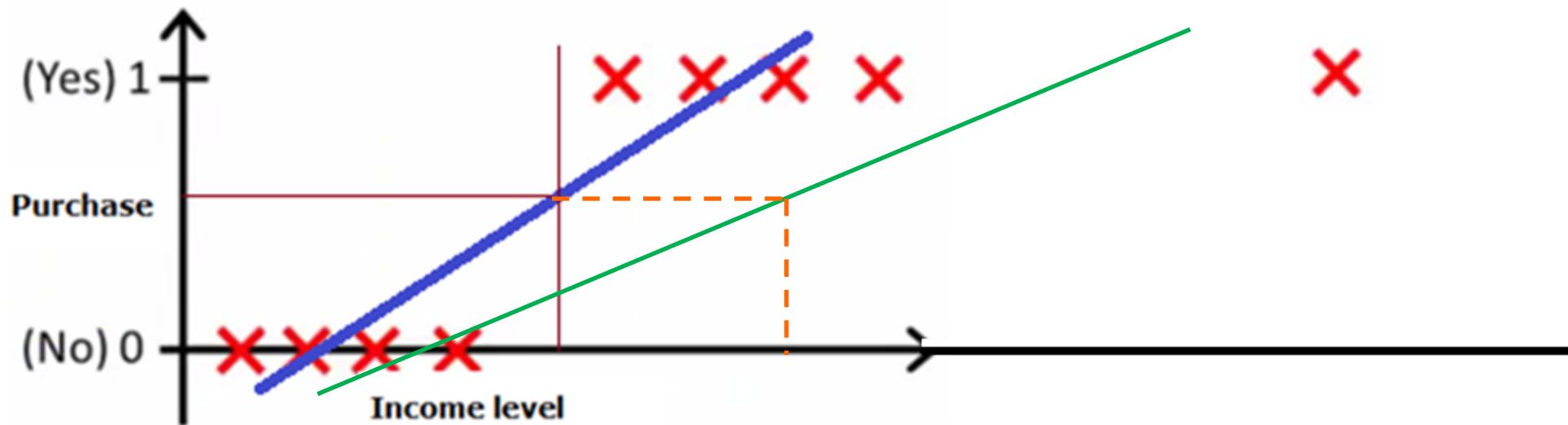


LOGISTIC REGRESSION

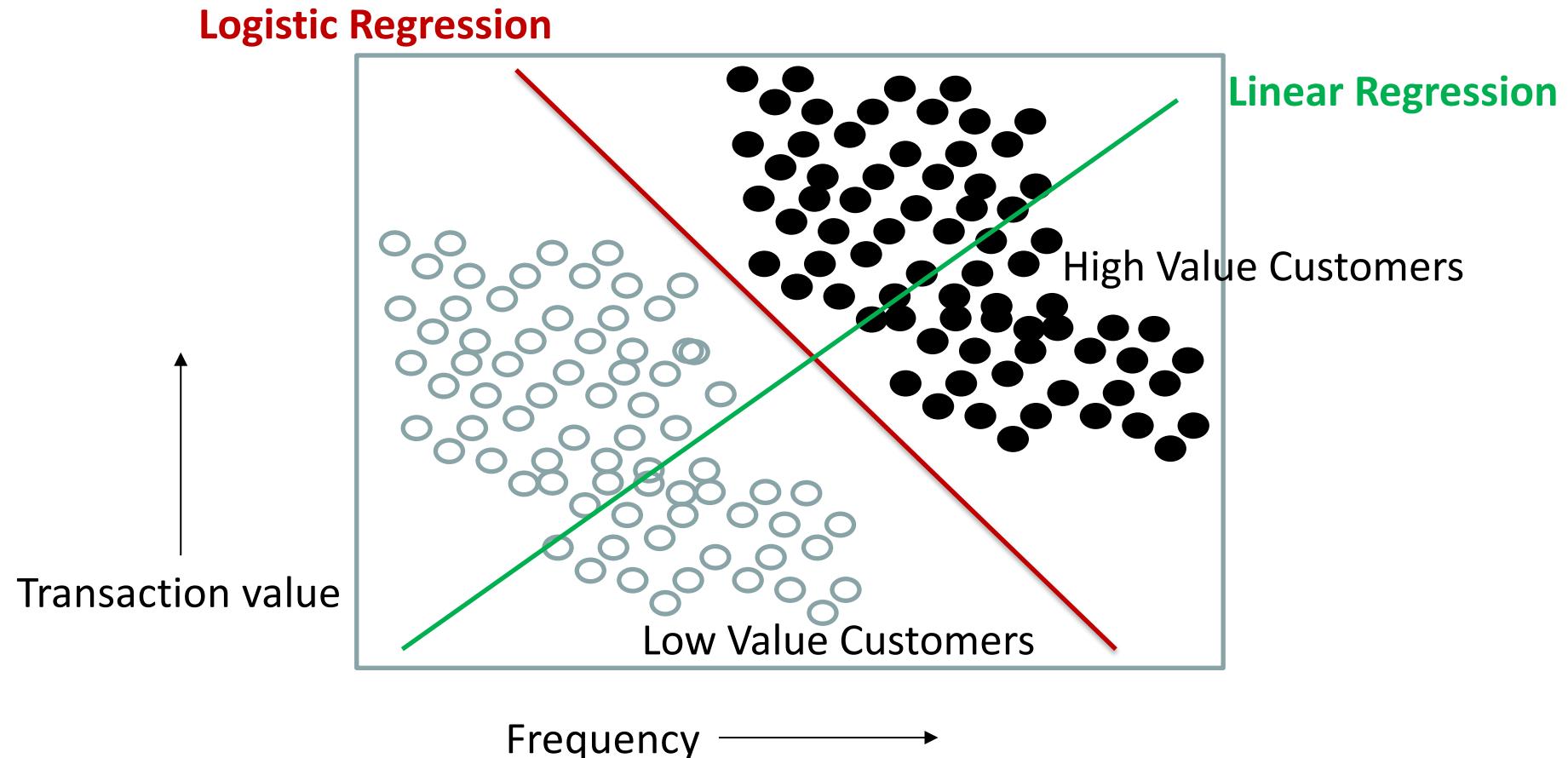
Classification Tasks: Regression



It could fail



Logistic Regression



In addition, linear regression outputs values in the entire range of $[-\infty, \infty]$, whereas the actual values in this case are bound by 0 and 1.

That is, $E(Y|X = x) = \hat{y} = \beta_0 + \beta_1 x_1 + \dots$ is not useful for such classification tasks. We need a function that can output values between 0 and 1.

A sigmoid or a logistic function allows that, and hence the name Logistic Regression.



Logistic Regression

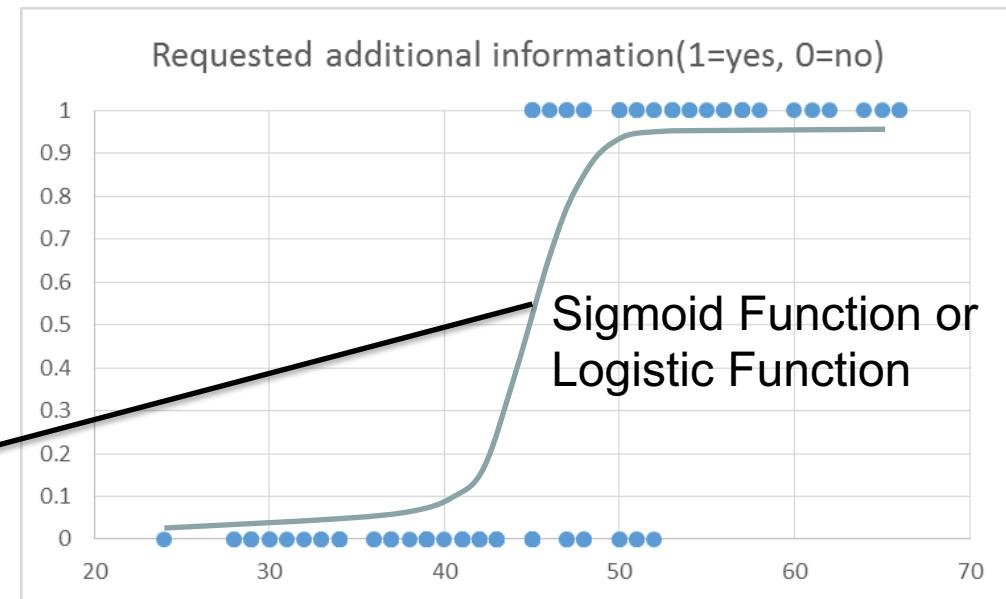
An auto club mails a flier to its members offering to send more information regarding a supplemental health insurance plan if the member returns a brief enclosed form.

Can a model be built to predict if a member will return the form or not?



Logistic model

$$E(Y|X = x) = f(x) = P(Y = 1|X = x) = \frac{1}{1 + e^{-\mu}} = \frac{e^\mu}{1 + e^\mu}$$



where $\mu = \beta_0 + \beta_1 x_1$ (also known as the systematic or the structural component or linear predictor).

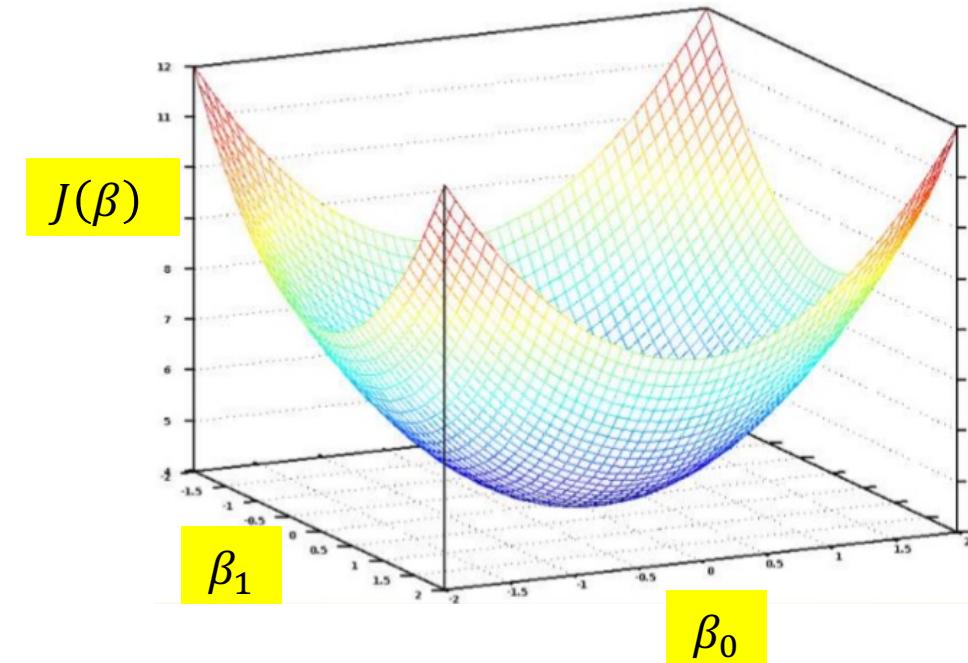
Note: For simplicity, we will use “p” to represent the above conditional probabilities going forward. Please keep in mind that “p” is a nonlinear function of β s.

This is a logistic model. We need to find the optimum β parameters.

Cost Function or Loss Function

In OLS Regression, parameters β were obtained so as to minimize the Cost Function or the Loss Function (squared errors in this case),

$$J(\beta) = \frac{\sum_{i=1, j=1}^{i=n, j=m} ((\beta_0 + \beta_j x_{ij}) - y_i)^2}{2n}$$
, where n is the # of observations and m are the number of independent variables.



$J(\beta)$ can be minimized using an Analytical/Calculus based approach, which gives the best linear unbiased estimators (β s) when the OLS assumptions are met.

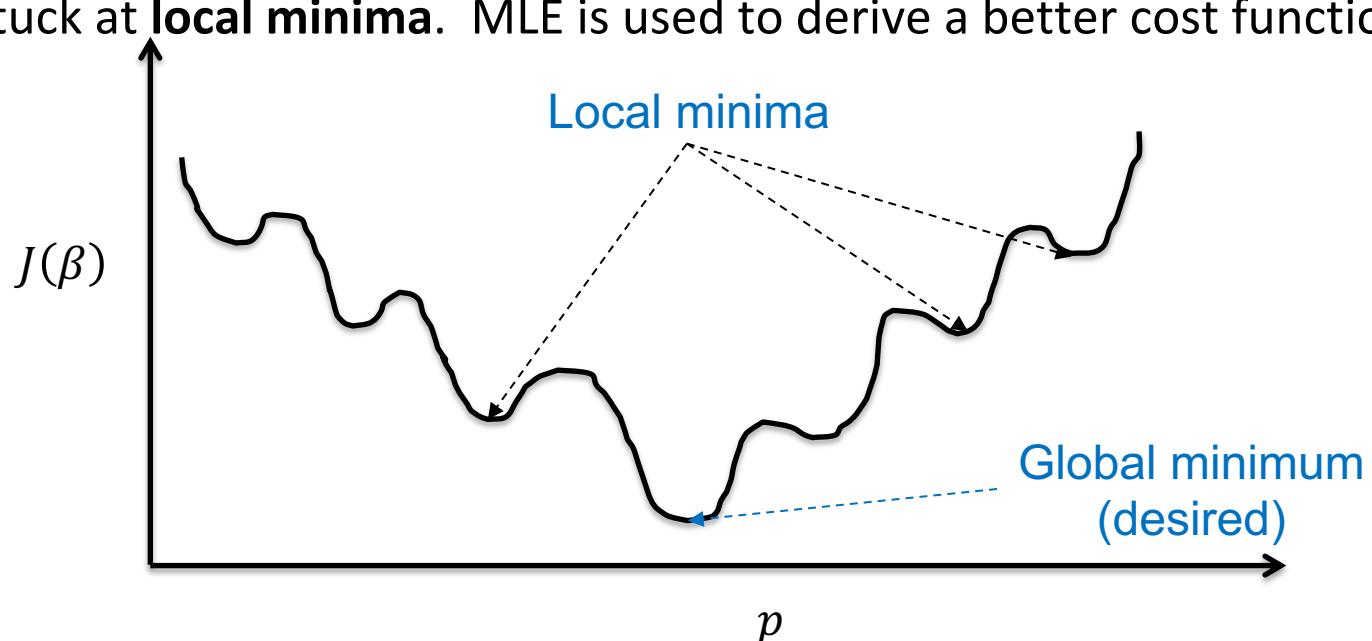
Image source: <https://www.slideshare.net/ocampesato/d3-typescript-and-deep-learning>
Last accessed: May 18, 2018

Maximum Likelihood Estimation (MLE) Principle to Derive Cost Function in Logistic Regression

A squared errors based cost function in Logistic Regression is highly non-linear leading to various minima because we are using a **non-linear sigmoid function** and then squaring it.

$$J(\beta) = \frac{\sum(p_i - y_i)^2}{2n}$$

The algorithm may get stuck at **local minima**. MLE is used to derive a better cost function.



Reference: Andrew Ng's Machine Learning course on Coursera

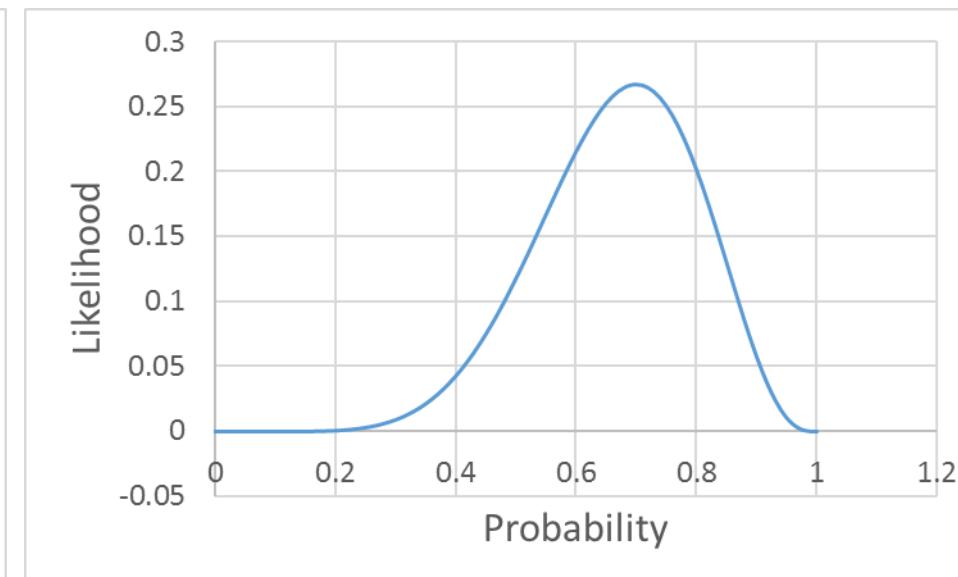
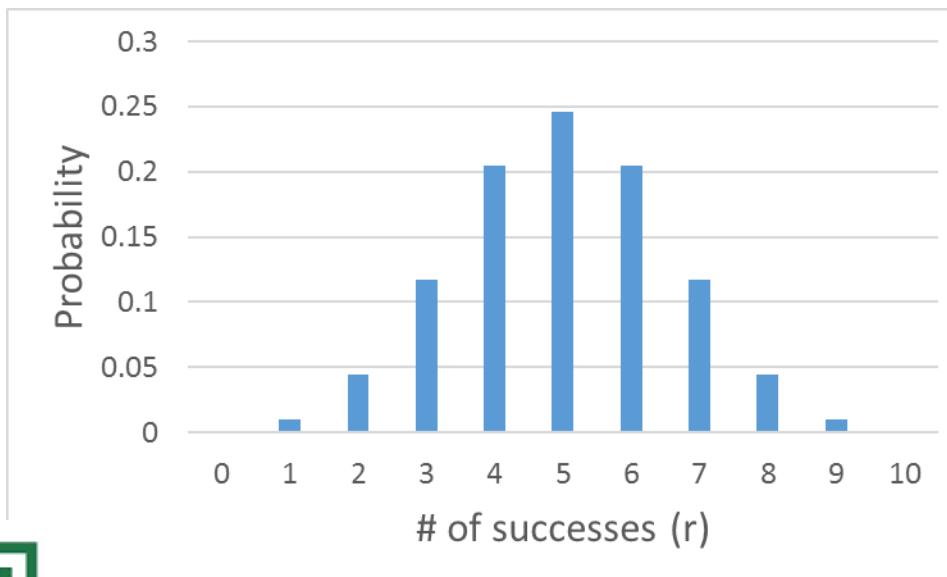


Maximum Likelihood Estimation (MLE)

INTUITION

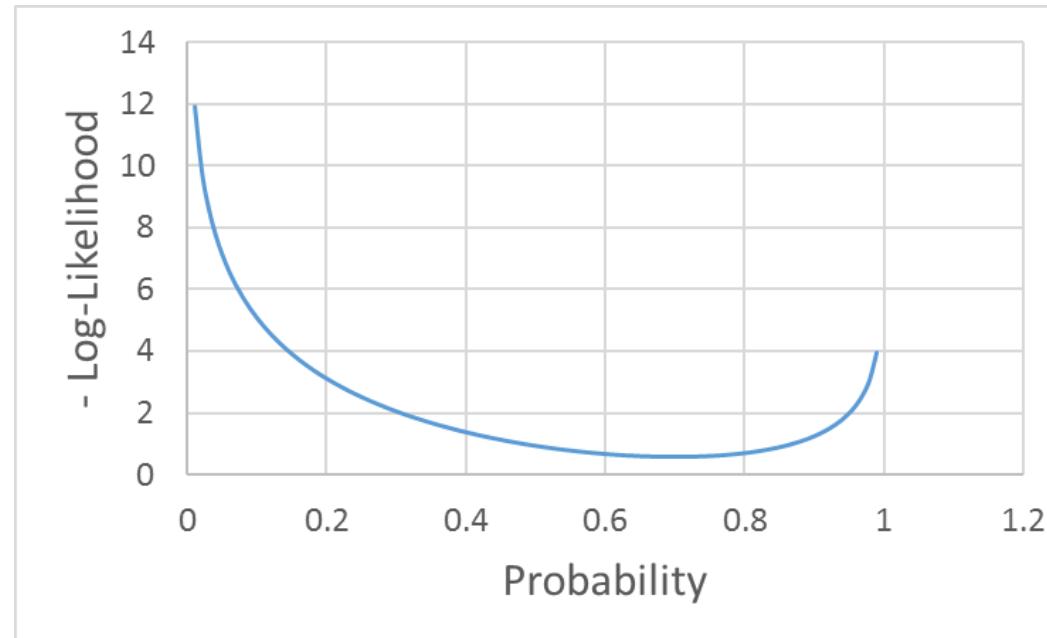
Probability vs Likelihood – Excel [“Probability –likelihood”]

- Likelihood is also known as reverse probability.
- In Probability, we **predict data** based on **known parameters**.
(Recall $B(n,p)$, $Geo(p)$, $Po(\lambda)$, $N(\mu,\sigma^2)$, etc.)
- In Likelihood, we **predict parameters** based on **known data**.



MLE

- Goal is to find parameters that **maximize the likelihood** of our observed data, which is the same as **minimizing errors**.
- We use calculus where functions are simple (minimize sum of squared errors in linear regression, and so on) or numerical techniques like Gradient Descent, Newton's Method, Fisher's Scoring, etc. (minimize deviance in logistic regression, Neural Nets, etc.), which work for both simple and complex functions, but become necessary for complex functions.
- So, Maximum Likelihood => Minimum of Negative Likelihood. That is, negative likelihood, or more specifically, **negative log-likelihood** is a measure of the errors of the model (what the model is unable to explain).



MLE to Derive Logistic Regression Objective and Cost Functions

We have understood that

$$L(\text{Parameters} | \text{Observations}) = P(\text{Observations} | \text{Parameters})$$

$$\therefore \log(L) = \log(p)$$

Remember, $P(Y=1 | X=x) = p$ and $P(Y=0 | X=x) = 1-p$

We can write $\log(p_{\text{class1}}) = 1 * \log(p) = y_i * \log(p)$ because $y_i = 1$ for class1.

Similarly, $\log(p_{\text{class2}}) = 1 * \log(1-p) = (1-y_i) * \log(1-p)$ because $y_i = 0$ for class2.

As either $y_i=0$ or $(1-y_i)=0$, we can combine the two to write

$$\log(p) = \log(L) = LL = y_i * \log(p) + (1-y_i) * \log(1-p)$$

This **LL** is the **Objective Function** we want to maximize. This is the same as minimizing the **-LL**, which becomes our **Cost Function**.

L: Likelihood
P: Probability
LL: Log-likelihood

Cost Function

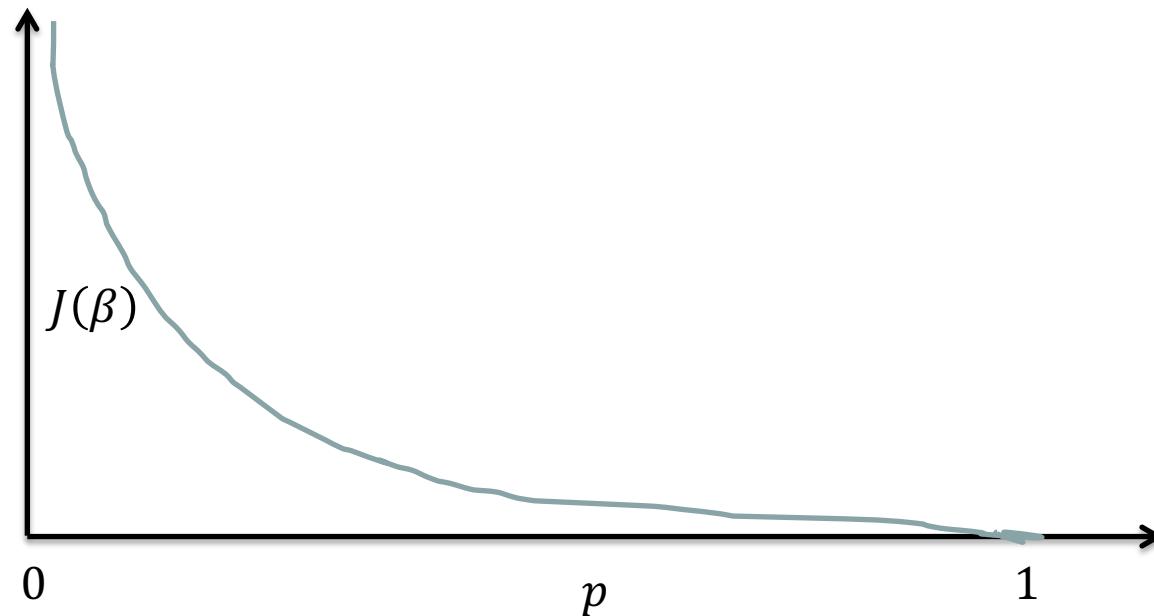
The cost function that provides a nice convex curve with a global minimum is given by:

$$J(\beta) = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(p) + (1 - y_i) \log(1 - p) \right]$$

Note: Don't forget p is the sigmoid function given by $P(Y = 1|X = x) = \frac{1}{1+e^{-\mu}} = \frac{e^\mu}{1+e^\mu}$, with μ being a linear function of β s.

What is the cost or penalty on the algorithm if $y_i = 1$, and the algorithm predicts with $p = 1$?

What is the cost or penalty on the algorithm if $y_i = 1$, and the algorithm predicts with $p = 0$?



Reference: Andrew Ng's Machine Learning course on Coursera

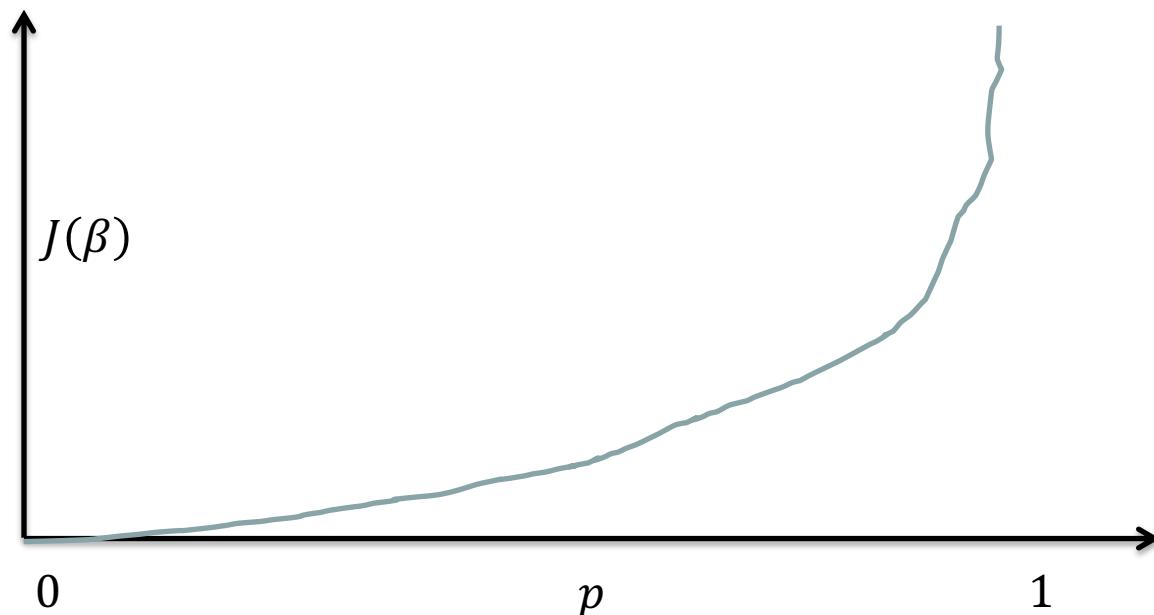
Cost Function

The cost function that provides a nice convex curve with a global minimum is given by:

$$J(\beta) = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(p) + (1 - y_i) \log(1 - p) \right]$$

What is the cost or penalty on the algorithm if $y_i = 0$, and the algorithm predicts with $p = 1$?

What is the cost or penalty on the algorithm if $y_i = 0$, and the algorithm predicts with $p = 0$?



Reference: Andrew Ng's Machine Learning course on Coursera

Cost Function

The cost function that provides a nice convex curve with a global minimum is given by:

$$J(\beta) = -\frac{1}{n} \left[\sum_{i=1}^n y_i \log(p) + (1 - y_i) \log(1 - p) \right]$$

There is no analytical solution to this, though. *Note p is a sigmoid function.* Algorithms like **Gradient Descent** are used instead.

Avoids assumptions regarding normality and homoscedasticity of errors, and linearity between dependent and independent variables.



Maxima and Minima – Gradient Descent

Maxima and Minima – Calculus Approach (Refresher)

$$y = x^3 - 2x^2 + x + 3$$

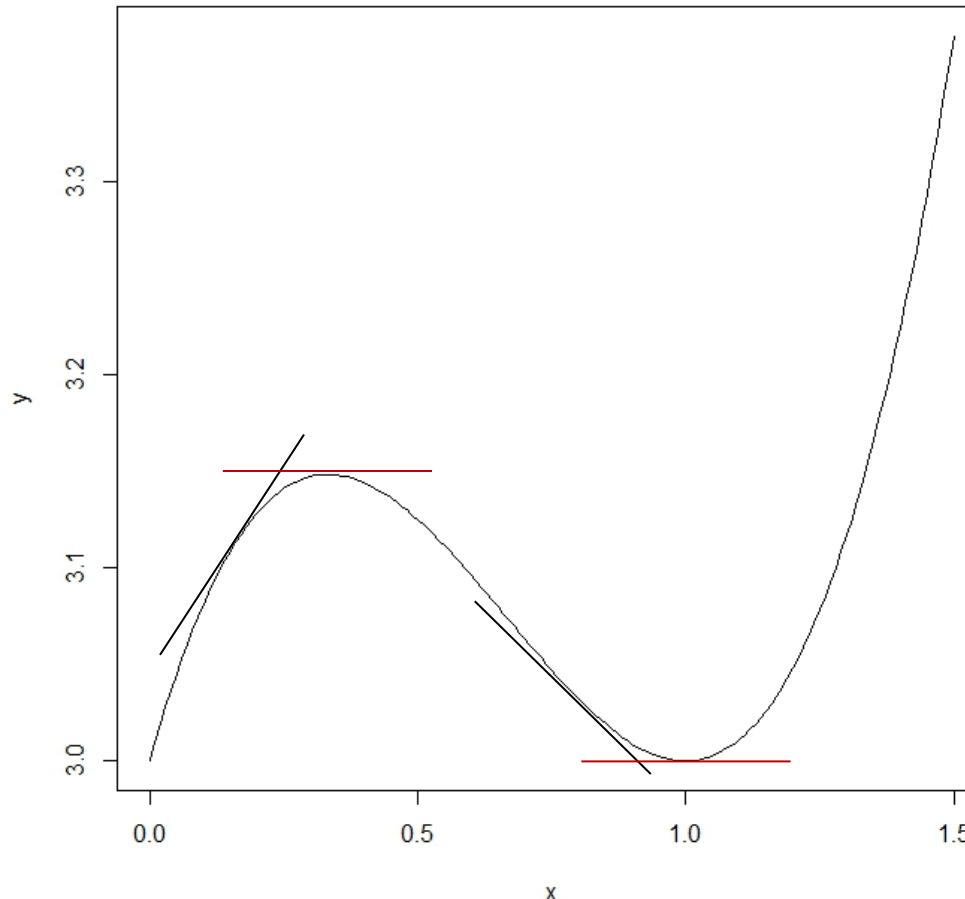
$$\frac{dy}{dx} = 3x^2 - 4x + 1 = 0$$

$$\frac{dy}{dx} = 0 \Rightarrow x = \frac{1}{3} \text{ or } 1$$

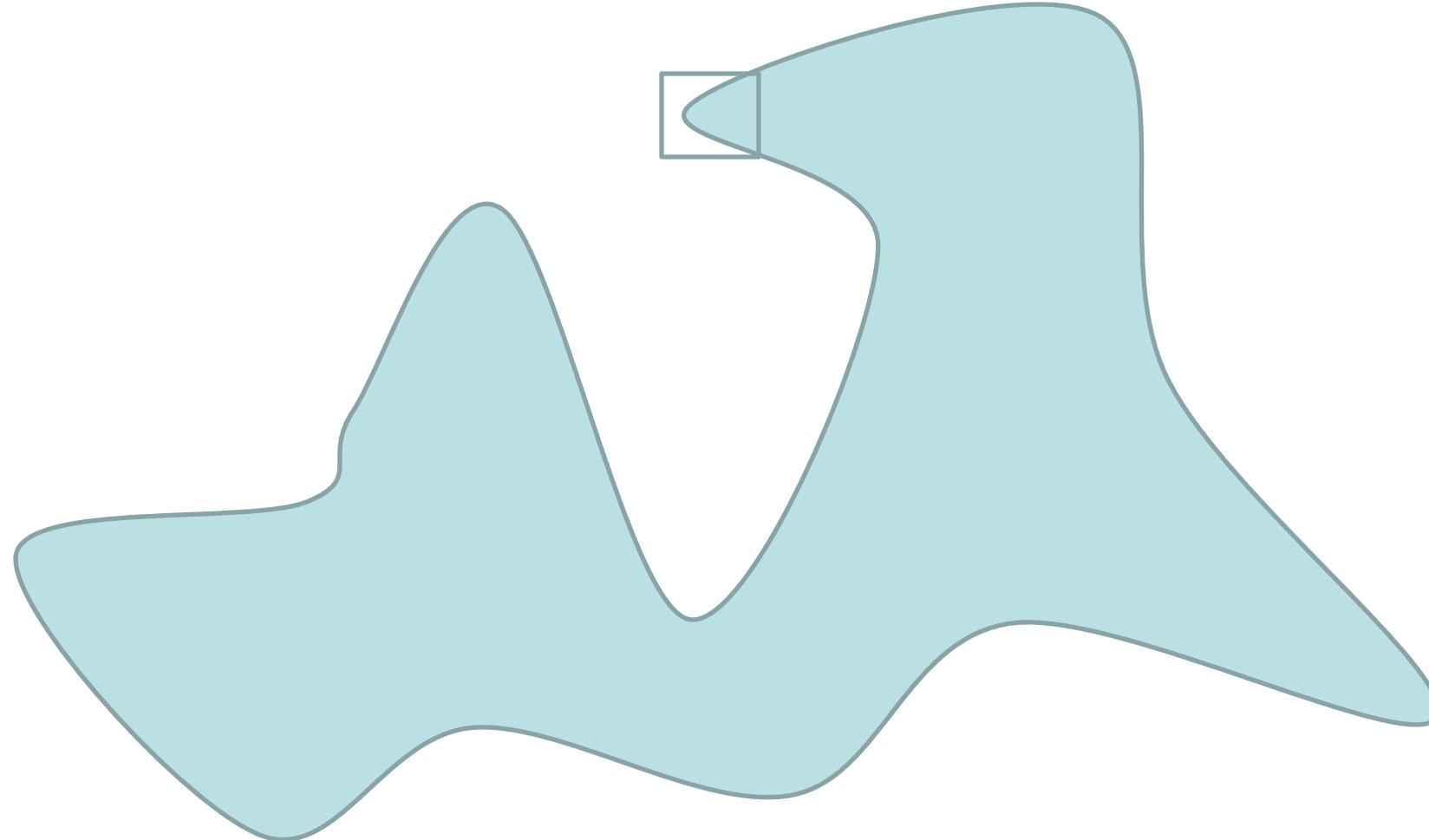
$$\frac{d}{dx} \left(\frac{dy}{dx} \right) = \frac{d^2y}{dx^2} = 6x - 4$$

for $x = 1$; $\frac{d^2y}{dx^2} = 2$. So, 1 is minima

for $x = \frac{1}{3}$; $\frac{d^2y}{dx^2} = -2$. So, $\frac{1}{3}$ is maxima



Finding Maxima and Minima For Extremely Complex Functions – Iterative Numerical Approach

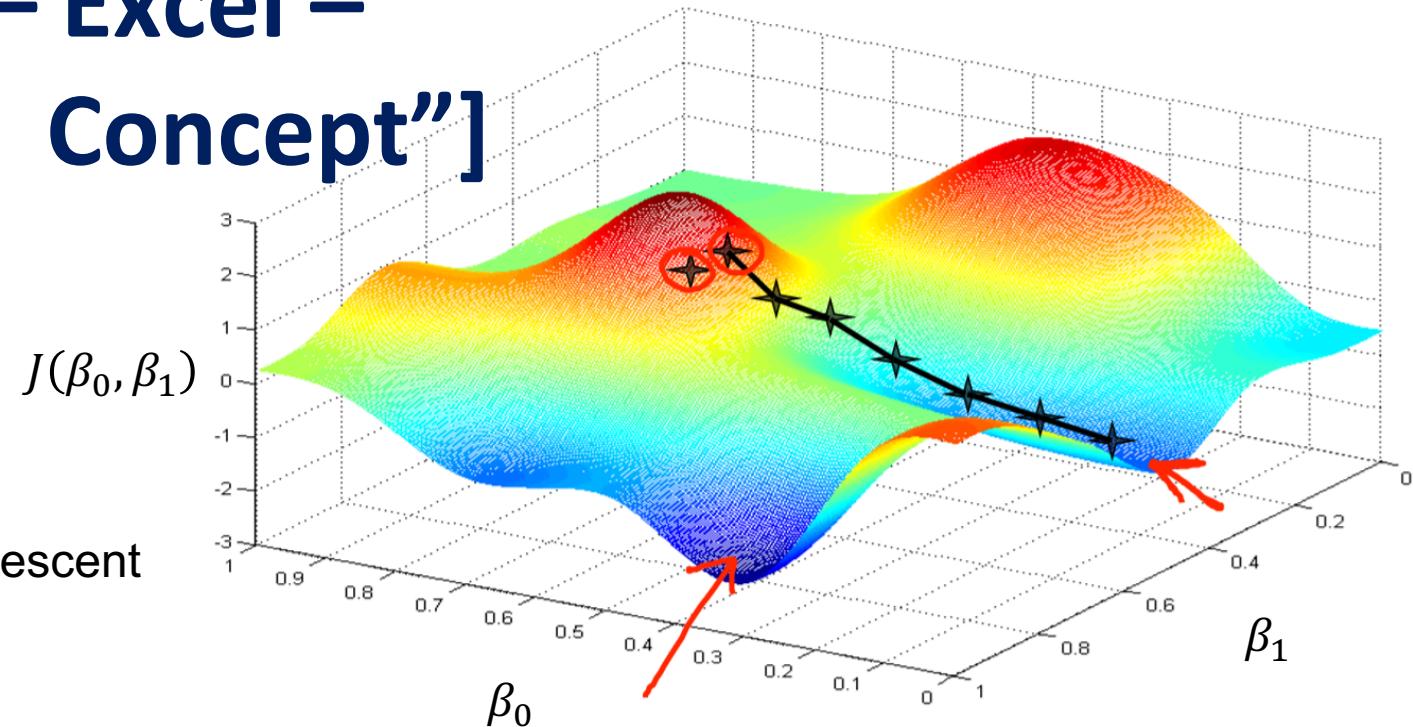


Gradient Descent – Excel –

[“Gradient Decent Concept”]

$$\beta_0^{(t+1)} = \beta_0 - \alpha \frac{\partial J(\beta_0, \beta_1)}{\partial \beta_0}$$

Learning Rate Slope of steepest descent



Source: Andrew Ng's Machine Learning course on Coursera

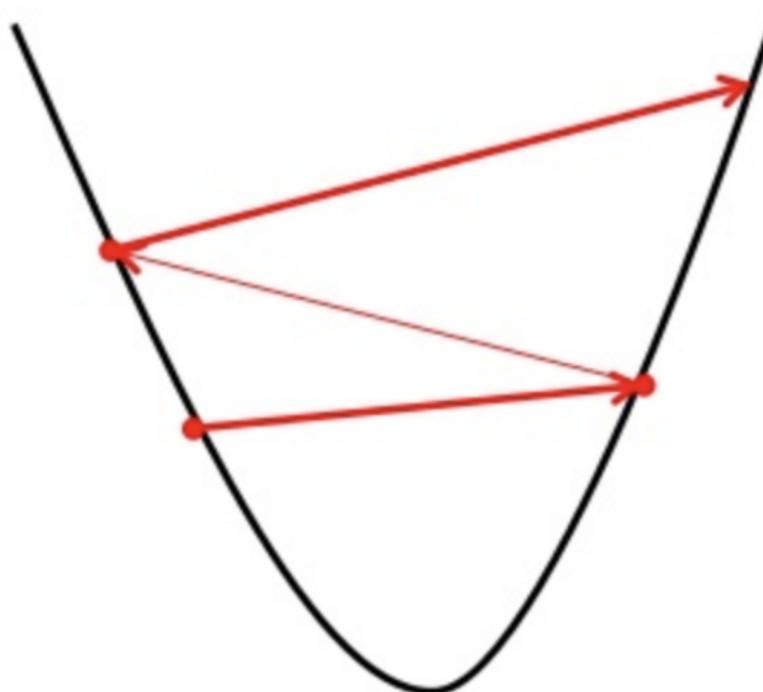
$$\beta_1^{(t+1)} = \beta_1 - \alpha \frac{\partial J(\beta_0, \beta_1)}{\partial \beta_1}$$



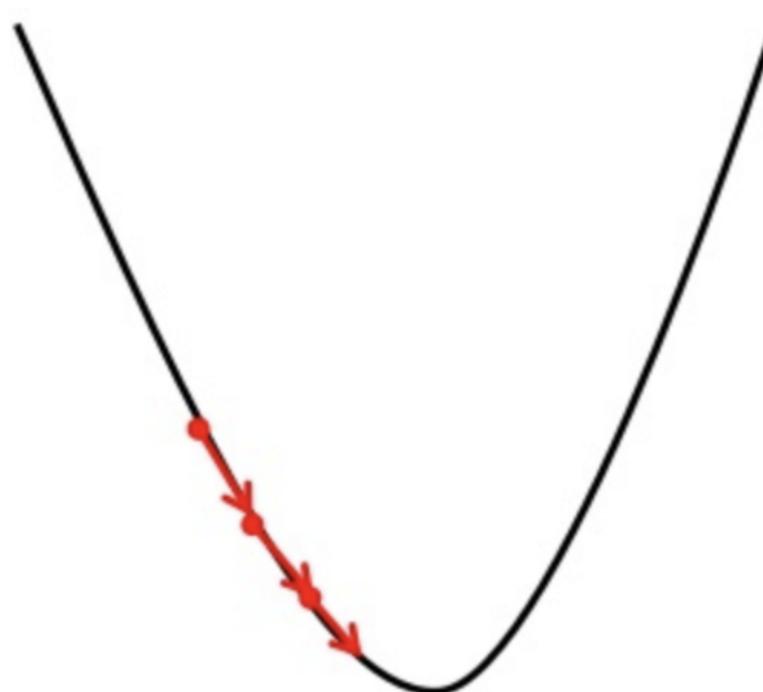
Learning Rate

BREAK

Big learning rate



Small learning rate



Source: <https://towardsdatascience.com/gradient-descent-in-a-nutshell-eaf8c18212f0>

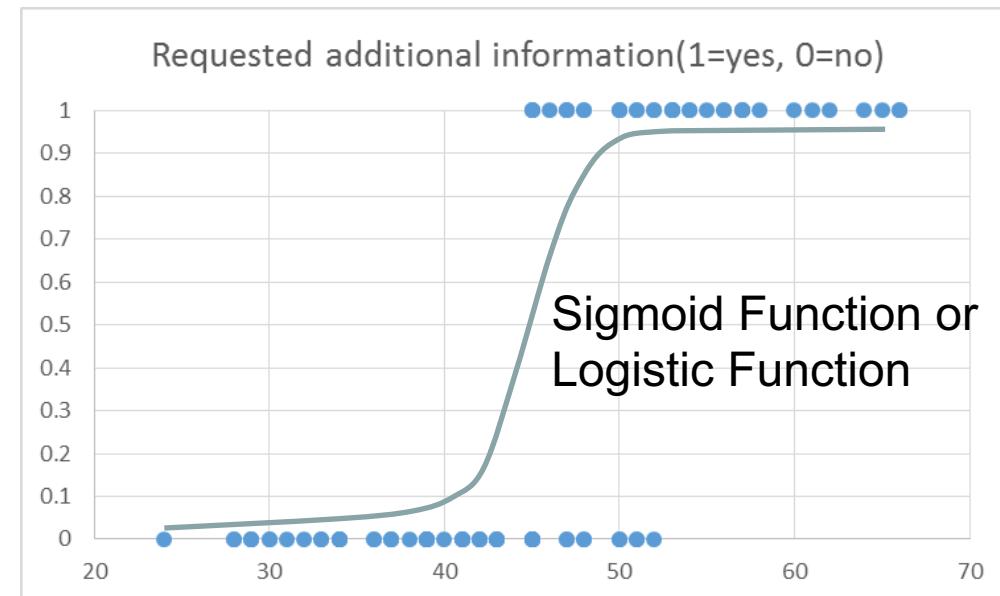
Logistic model

Odds are obtained by the probability of an event occurring divided by the probability that it will not occur.

Logistic model can be transformed into Odds:

$$S = Odds = \frac{p}{1 - p}$$

$$S = Odds = \frac{Success}{Failure}$$



$$f(x) = p = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}$$

Attention Check – Probability and Odds

If the probability of winning is 6/12, what are the odds of winning?	1:1 (Note, the probability of losing also is 6/12)
If the odds of winning are 13:2, what is the probability of winning?	13/15
If the odds of winning are 3:8, what is the probability of losing?	8/11
If the probability of losing is 6/8, what are the odds of winning?	2:6 or 1:3

$$\text{Odds of winning} = \frac{\text{Probability of Winning (Success)}}{\text{Probability of Losing (Failure)}}$$

Logistic model

$$S = Odds = \frac{p}{1 - p}$$

$$S = \frac{\frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}}{1 - \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}} \quad \therefore S = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}$$

$$\ln(S) = \ln(e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Logistic model

The log of the odds is called logit, and the transformed model is linear in β s.

Solving the Logistic regression problem essentially reduces to finding the set of β s that minimizes error.





R and Interpreting the output

```
call:  
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.95015	-0.32016	-0.05335	0.26538	1.72940

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-20.40782	4.52332	-4.512	6.43e-06 ***
Age	0.42592	0.09482	4.492	7.05e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 123.156 on 91 degrees of freedom  
Residual deviance: 49.937 on 90 degrees of freedom  
AIC: 53.937
```

Number of Fisher Scoring iterations: 7

What is the logit equation?

$$\ln(S) = -20.40782 + 0.42592Age$$

Determining Logistic Regression Model

Suppose we want a probability that a 50-year old club member will return the form.

$$\ln(S) = -20.40782 + 0.42592 * 50 = 0.89$$

$$S = e^{0.89} = 2.435$$

The odds that a 50-year old returns the form are 2.435 to 1.



Determining Logistic Regression Model

$$\hat{p} = \frac{S}{S + 1} = \frac{2.435}{2.435 + 1} = 0.709$$

[Remember $S = Odds = \frac{p}{1-p}$]

Using a probability of 0.50 as a cutoff between predicting a 0 or a 1, this member would be classified as a 1.

The output of the logistic regression is a probability value. You need to fix a threshold value before a class is assigned.

Computing using R

What is the probability that a 50 year-old will return the form?

```
> flierresponseglm <- glm(Response~Age, data = flierresponse, family = "binomial")
> nd <- data.frame(Age=50) #To predict the probability for Age=50, put that info in a data-frame
> predict(flierresponseglm,newdata=nd) # This gives the log-Odds
   1
0.8879707
> predict(flierresponseglm,newdata=nd,type="response") # Compute the probability
   1
0.7084712
```

Interpreting Output - Deviances

Deviance or Residual Deviance is *similar to SSE* in the sense it measures how much remains unexplained by the model built with predictors included.

$$D = -2LL = -2 * \left[\sum_{i=1}^n y_i \log(p) + (1 - y_i) \log(1 - p) \right]$$

where LL is the log-likelihood.

Null Deviance shows how well the model predicts the response with only the intercept as a parameter. The intercept is the logarithm of the ratio of cases with $y=1$ to the number of cases with $y=0$. This is *similar to SST*, which gives total variation when all coefficients are zero (null hypothesis).

```
Call:  
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)  
  
Deviance Residuals:  
    Min      1Q   Median      3Q     Max  
-1.95015 -0.32016 -0.05335  0.26538  1.72940  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -20.40782   4.52332 -4.512 6.43e-06 ***  
Age          0.42592   0.09482  4.492 7.05e-06 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 123.156 on 91 degrees of freedom  
Residual deviance: 49.937 on 90 degrees of freedom  
AIC: 53.937  
  
Number of Fisher Scoring iterations: 7
```



Interpreting Output – Testing the Overall Model

The z-values and the associated *p*-values provide significance of individual predictor variables.

```
call:  
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-1.95015 -0.32016 -0.05335  0.26538  1.72940  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -20.40782   4.52332 -4.512 6.43e-06 ***  
Age          0.42592   0.09482  4.492 7.05e-06 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 123.156 on 91 degrees of freedom  
Residual deviance: 49.937 on 90 degrees of freedom  
AIC: 53.937  
  
Number of Fisher Scoring iterations: 7
```

Interpreting Output – Testing the Overall Model

- AIC provides a means for model selection.
- $AIC = D + 2k$, where k is the # of parameters in the model including the intercept. Recall in Linear Regression, it is calculated as $AIC = n\ln(RSS/n) + 2k$.
- AIC is *similar to Adjusted R²* in the sense it penalizes for adding more parameters to the model.



Applications

- Predicting stock price movement (up/down)
- Predict whether a patient has diabetes or not
- Predict whether a customer will buy or not
- Predict the likelihood of loan default



Diagnostic Hints

- Overly large coefficient magnitudes, overly large error bars on the coefficient estimates, and the wrong sign on a coefficient could be indications of correlated inputs.
- VIF can be used to check for multicollinearity. R outputs a Generalized Variance Inflation Factor, which is obtained by correcting VIF to the degrees of freedom for categorical predictors. $GVIF = VIF^{\left(\frac{1}{2*df}\right)}$

Case – Framingham Heart Study



Framingham Heart Study

A Project of the National Heart, Lung, and Blood Institute and Boston University

- Committed to identifying common factors contributing to cardiovascular disease (CVD).
- Setup in the town of Framingham, MA in 1948.
- Random sample consisting of 2/3rds of adult population in the town.

AGE-SEX DISTRIBUTION AT ENTRY (1948)				
Age	29-39	40-49	50-62	Totals
Men	835	779	722	2,336
Women	1,042	962	869	2,873
Totals	1,877	1,741	1,591	5,209

Case Study – Data (framinghamheartstudy.org and MITx)

- 5209 men and women participated.
- Age range: 30-62
- People who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke.
- Careful monitoring of Framingham Study population has led to identification of major CVD risk factors.
- Led to development of Framingham Risk Score, a gender specific algorithm used to estimate the 10-year cardiovascular risk of an individual:
<http://cvdrisk.nhlbi.nih.gov/>



Case Study – Data (framinghamheartstudy.org and MITx)

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	1	39	4	0	0	0	0	0	0	195	106.0	70.0	26.97	80	77	0
2	0	46	2	0	0	0	0	0	0	250	121.0	81.0	28.73	95	76	0
3	1	48	1	1	20	0	0	0	0	245	127.5	80.0	25.34	75	70	0
4	0	61	3	1	30	0	0	1	0	225	150.0	95.0	28.58	65	103	1
5	0	46	3	1	23	0	0	0	0	285	130.0	84.0	23.10	85	85	0
6	0	43	2	0	0	0	0	1	0	228	180.0	110.0	30.30	77	99	0
7	0	63	1	0	0	0	0	0	0	205	138.0	71.0	33.11	60	85	1
8	0	45	2	1	20	0	0	0	0	313	100.0	71.0	21.68	79	78	0
9	1	52	1	0	0	0	0	1	0	260	141.5	89.0	26.36	76	79	0
10	1	43	1	1	30	0	0	1	0	225	162.0	107.0	23.61	93	88	0
11	0	50	1	0	0	0	0	0	0	254	133.0	76.0	22.91	75	76	0
12	0	43	2	0	0	0	0	0	0	247	131.0	88.0	27.64	72	61	0
13	1	46	1	1	15	0	0	1	0	294	142.0	94.0	26.31	98	64	0
14	0	41	3	0	0	1	0	1	0	332	124.0	88.0	31.31	65	84	0
15	0	39	2	1	9	0	0	0	0	226	114.0	64.0	22.35	85	NA	0
16	0	38	2	1	20	0	0	1	0	221	140.0	90.0	21.35	95	70	1
17	1	48	3	1	10	0	0	1	0	232	138.0	90.0	22.37	64	72	0
18	0	46	2	1	20	0	0	0	0	291	112.0	78.0	23.38	80	89	1
19	0	38	2	1	5	0	0	0	0	195	122.0	84.5	23.24	75	78	0
20	1	41	2	0	0	0	0	0	0	195	139.0	88.0	26.88	85	65	0
21	0	42	2	1	30	0	0	0	0	190	108.0	70.5	21.59	72	85	0
22	0	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0
23	0	52	1	0	0	0	0	0	0	234	148.0	78.0	34.17	70	113	0
24	0	52	3	1	20	0	0	0	0	215	132.0	82.0	25.11	71	75	0
25	1	44	2	1	30	0	0	1	0	270	137.5	90.0	21.96	75	83	0
26	1	47	4	1	20	0	0	0	0	294	102.0	68.0	24.18	62	66	1

Case Study – Predicting Coronary Heart Disease (CHD)

Data description

4240 observations; 15 predictor and 1 predicted variables

- *TenYearCHD* – To be predicted. Risk of having a heart attack or stroke in the next 10 years.

Predictors

- Demographic Risk Factors
 - *male*: Gender of subject – Yes or No
 - *age*: Age of subject at first examination
 - *education*: some high school (1), high school (2), some college/vocational college (3), college (4)

Case Study – Predicting Coronary Heart Disease (CHD)

- Behavioural Risk Factors
 - *currentSmoker*: Yes or No
 - *cigsPerDay*: No. of cigarettes smoked per day if smoker
- Medical History Risk Factors
 - *BPmeds*: On BP medication at the time of first examination – Yes or No
 - *prevStroke*: Did the subject have a previous stroke – Yes or No
 - *prevHyp*: Is the subject currently hypertensive – Yes or No
 - *diabetes*: Does the subject currently have diabetes – Yes or No



Case Study – Predicting Coronary Heart Disease (CHD)

- Risk Factors from First Examination
 - *totChol*: Total cholesterol (mg/dL)
 - *sysBP*: Systolic blood pressure (the higher number in BP result)
 - *diaBP*: Diastolic blood pressure (the lower number in BP result)
 - *BMI*: Body Mass Index (kg/m^2)
 - *heartRate*: # of beats per minute
 - *glucose*: Blood glucose level (mg/dL)

Case Study – Predicting Coronary Heart Disease (CHD)

Approach

- Randomly split data into training and test in 70:30 ratio.
- Measure prediction accuracies on training and test data
- Although the split is random, make sure the proportions of the categories are roughly the same in both training and test sets.

```
# Randomly split the data into training and testing sets
set.seed(1000)
split = sample.split(framingham$TenYearCHD, SplitRatio = 0.70)

# Split up the data using subset
train = subset(framingham, split==TRUE)
test = subset(framingham, split==FALSE)

# Check the proportions of CHD in both sets
cat(sum(train$TenYearCHD)/nrow(train),sum(test$TenYearCHD)/nrow(test))
0.1519542    0.1517296
```



Case Study – Predicting Coronary Heart Disease (CHD)

Results

- Significant variables that cannot be controlled
 - Gender
 - Age
 - Medical history
- Significant variables that can be controlled
 - Smoking habits
 - Cholesterol
 - Systolic BP
 - Blood glucose

```
call:  
glm(formula = TenYearCHD ~ ., family = binomial, data = train)  
  
Deviance Residuals:  
    Min      1Q   Median      3Q     Max  
-1.9392 -0.5998 -0.4211 -0.2771  2.8632  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -8.360272  0.864696 -9.668 < 2e-16 ***  
male          0.524080  0.130836  4.006 6.19e-05 ***  
age           0.065429  0.008049  8.129 4.34e-16 ***  
education     -0.041105  0.059185 -0.695 0.487366  
currentSmoker  0.120498  0.187629  0.642 0.520735  
cigsPerDay    0.016471  0.007488  2.200 0.027825 *  
BPMedS        0.169118  0.282140  0.599 0.548898  
prevalentstroke 1.156666  0.560179  2.065 0.038940 *  
prevalentHyp   0.307077  0.166034  1.849 0.064389 .  
diabetes       -0.319937  0.392574 -0.815 0.415087  
totChol        0.003799  0.001330  2.856 0.004290 **  
sysBP          0.011144  0.004446  2.507 0.012188 *  
diaBP          -0.001861  0.007760 -0.240 0.810517  
BMI            0.008812  0.015662  0.563 0.573702  
heartRate      -0.007273  0.005131 -1.418 0.156296  
glucose         0.009227  0.002752  3.353 0.000798 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 2176.6 on 2565 degrees of freedom  
Residual deviance: 1919.9 on 2550 degrees of freedom  
(402 observations deleted due to missingness)  
AIC: 1951.9
```

Missing Values

There are several ways of dealing with missing values.

If large percentage of data for a given variable is missing, then we don't use that variable for building the model.

If the percentage of missing values is small (5 to 10%)

- Naïve method: Replace the missing values with either mean, median or mode
- Intelligent method: Impute the missing values from the relationship between the variables

Also see: <https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/>



VIF, GVIF, GVIF $\left(\frac{1}{2*df}\right)$

Predicting Coronary Heart Disease Case

```
call:  
glm(formula = TenYearCHD ~ ., family = binomial, data = train)  
  
Deviance Residuals:  
    Min      1Q   Median      3Q     Max  
-1.9392 -0.5998 -0.4211 -0.2771  2.8632  
  
Coefficients:  
             Estimate Std. Error z value Pr(>|z|)  
(Intercept) -8.360272  0.864696 -9.668 < 2e-16 ***  
male          0.524080  0.130836  4.006 6.19e-05 ***  
age           0.065429  0.008049  8.129 4.34e-16 ***  
education     -0.041105  0.059185 -0.695 0.487366  
currentSmoker 0.120498  0.187629  0.642 0.520735  
cigsPerDay    0.016471  0.007488  2.200 0.027825 *  
BPMeds        0.169118  0.282140  0.599 0.548898  
prevalentstroke 1.156666  0.560179  2.065 0.038940 *  
prevalentHyp   0.307077  0.166034  1.849 0.064389 .  
diabetes       -0.319937  0.392574 -0.815 0.415087  
totchol        0.003799  0.001330  2.856 0.004290 **  
sysBP          0.011144  0.004446  2.507 0.012188 *  
diabP          -0.001861  0.007760 -0.240 0.810517  
BMI            0.008812  0.015662  0.563 0.573702  
heartRate      -0.007273  0.005131 -1.418 0.156296  
glucose        0.009227  0.002752  3.353 0.000798 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 2176.6 on 2565 degrees of freedom  
Residual deviance: 1919.9 on 2550 degrees of freedom  
(402 observations deleted due to missingness)  
AIC: 1951.9
```

- Education as **numeric** variable

```
> framingham = read.csv("framingham.csv")  
> str(framingham)  
'data.frame': 4240 obs. of 16 variables:  
 $ male      : int  1 0 1 0 0 0 0 1 1 ...  
 $ age       : int  39 46 48 61 46 43 63 45 52 43 ...  
 $ education : int  4 2 1 3 3 2 1 2 1 1 ...
```

- car package gives the following VIF figures

```
> car::vif(framinghamLog)  
               male              age            education  
             1.247670           1.278996          1.057810
```

VIF, GVIF, GVIF $\left(\frac{1}{2*df}\right)$

Predicting Coronary Heart Disease Case

```

Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = train)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.9315 -0.5948 -0.4196 -0.2733  2.8925 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -8.286327  0.856000 -9.680 < 2e-16 ***
male         0.506116  0.131991  3.834 0.000126 *** 
age          0.064020  0.008135  7.869 3.56e-15 *** 
education2  -0.210890  0.148310 -1.422 0.155038  
education3  -0.120464  0.174595 -0.690 0.490220  
education4  -0.082216  0.199469 -0.412 0.680212  
currentSmoker 0.125147  0.187693  0.667 0.504921  
cigsPerDay   0.016589  0.007482  2.217 0.026611 *  
BPMeds       0.177341  0.282452  0.628 0.530094  
prevalentStroke 1.158996  0.563826  2.056 0.039822 * 
prevalentHyp   0.308709  0.166304  1.856 0.063412 .  
diabetes      -0.318608  0.393231 -0.810 0.417807  
totChol       0.003860  0.001334  2.894 0.003801 ** 
sysBP         0.011195  0.004451  2.515 0.011893 * 
diaBP         -0.001726  0.007766 -0.222 0.824120  
BMI           0.007535  0.015680  0.481 0.630828  
heartRate     -0.007132  0.005136 -1.389 0.164946  
glucose       0.009221  0.002748  3.356 0.000791 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2176.6 on 2565 degrees of freedom
Residual deviance: 1918.2 on 2548 degrees of freedom
(402 observations deleted due to missingness)
AIC: 1954.2

Number of Fisher Scoring iterations: 5

```

- Education as **categorical** variable

```

> framingham$education = factor(framingham$education)
> str(framingham)
'data.frame':  4240 obs. of  16 variables:
 $ male        : int  1 0 1 0 0 0 0 1 1 ...
 $ age         : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education   : Factor w/ 4 levels "1","2","3","4": 4 2
 1 3 3 2 1 2 1 1 ...

```

- car package gives the following (G)VIF figures

```

> car::vif(framinghamLog)
              GVIF Df GVIF^(1/(2*Df))
male           1.268872  1      1.126442
age            1.302896  1      1.141445
education      1.121533  3      1.019300

```

- Use the **square** of the GVIF $\left(\frac{1}{2*df}\right)$ value and apply the VIF rule of thumb



A slight detour



HOW GOOD IS YOUR CLASSIFICATION?

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

		Predicted		METRICS
		Positive	Negative	
Actual	Positive	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	Negative	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
Precision				Accuracy, F_1 score

Confusion Matrix - Metrics

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	Negative	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{Actual+ve}}$$

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{True+ve} + \text{False-ve}}$$

$$\text{Specificity} = \frac{\text{True -ve}}{\text{Actual-ve}}$$

$$\text{Specificity} = \frac{\text{True -ve}}{\text{False+ve} + \text{True -ve}}$$

$$\text{Precision} = \frac{\text{True +ve}}{\text{Predicted +ve}}$$

$$\text{Precision} = \frac{\text{True+ve}}{\text{True+ve} + \text{False+ve}}$$

$$\text{Accuracy} = \frac{\text{True +ve} + \text{True-ve}}{\text{Total}}$$

$$\text{Accuracy} = \frac{\text{True+ve} + \text{True -ve}}{\text{True+ve} + \text{False-ve} + \text{False+ve} + \text{True -ve}}$$

$$F_1 \text{ Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion Matrix

Spam filtering		Predicted		Total	
		Positive	Negative		
Actual	Positive	952	526	1478	Recall(Sensitivity)
	Negative	167	3025	3192	Specificity
Total		1119	3551	4670	
		Precision			Accuracy, F1 Score

$$\text{Recall (Sensitivity)} = \frac{952}{1478} = 0.644$$

$$\text{Specificity} = \frac{3025}{3025 + 167} = \frac{3025}{3192} = 0.948$$

$$\text{Precision} = \frac{952}{1119} = 0.851$$

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = \frac{1.096}{1.495} = 0.733$$

Which measure(s) is/are more important?

$$\text{Accuracy} = \frac{952 + 3025}{952 + 3025 + 526 + 167} = \frac{3977}{4670} = 0.852$$

Confusion Matrix

Court System – Death Sentence		Verdict		
		Guilty	Not Guilty	
Actual	Guilty	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Not Guilty	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Which measure(s) is/are more important?



Confusion Matrix

Breast cancer detection		Predicted		Total
		Positive	Negative	
Actual	Positive	852	126	978
	Negative	67	1025	1092
Total		919	1151	2070

$$\text{Recall (Sensitivity)} = \frac{852}{978} = 0.871$$

$$\text{Precision} = \frac{852}{919} = 0.927$$

$$\text{Accuracy} = \frac{852 + 1025}{852 + 1025 + 126 + 67} = \frac{1877}{2070} = 0.907$$

$$\text{Specificity} = \frac{1025}{1025 + 67} = \frac{1025}{1092} = 0.939$$

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * 0.871 * 0.927}{0.871 + 0.927} = \frac{1.615}{1.798} = 0.898$$

Which measure(s) is/are more important?

Confusion Matrix

Anti Virus Detection		Detection		
		Virus	No Virus	
Actual	Virus	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	No Virus	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Which measure(s) is/are more important?

Confusion Matrix

Organ Matching from Donors		Predicted		
		Match	No Match	
Actual	Match	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	No Match	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Which measure(s) is/are more important?



Confusion Matrix

Credit Card Fraud Detection		Detection		
		Fraud	No Fraud	
Actual	Fraud	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	No Fraud	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Which measure(s) is/are more important?



Confusion Matrix

Image Text Classification		Predicted Word		
		CAT	DOG	
Actual Word	CAT	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	DOG	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Which measure(s) is/are more important?

Case Study – Predicting Coronary Heart Disease (CHD)

Results

- Accuracy in training set = $2200/2566 = 85.7\%$
- Accuracy in testing set = $927/1092 = 84.9\%$
- Accuracy is affected by **imbalance** between positives and negatives.
- There is a **trade-off** between sensitivity and specificity.

$$\text{Accuracy} = \frac{\text{True +ve} + \text{True -ve}}{\text{Total}}$$

$$\text{Accuracy} = \frac{\text{True+ve} + \text{True -ve}}{\text{True+ve} + \text{False-ve} + \text{False+ve} + \text{True -ve}}$$

Training Set

10-year CHD risk		Predicted	
Actual		True	False
	True	30	357
	False	9	2170

Testing Set

10-year CHD risk		Predicted	
Actual		True	False
	True	13	157
	False	6	916





Cohen's Kappa, ROC Curve, Gains and Lift Charts

SOME MORE PERFORMANCE METRICS

Cohen's Kappa Metric

- Accuracy can often be a misleading metric when one category occurs more often than other in the given data-set. For example:
 - Occurrence of cancer in general population is 0.4%
 - If a prediction system blindly marks everyone as “No cancer”, it will be 99.6% accurate



Cohen's Kappa Metric

- Kappa metric compares **Observed Accuracy** with **Expected Accuracy** (by random chance).

$$\kappa = \frac{\text{Observed Accuracy} - \text{Expected Accuracy}}{1 - \text{Expected Accuracy}}$$

$$\text{Observed Accuracy} = \frac{\text{Correct Predictions (True Positives+True Negatives)}}{\text{Total}}$$

Expected (Random chance) Accuracy uses Expected Frequencies, which are calculated the same way we did in Chi-Square calculation (Recall *Expected Frequency* = $\frac{\text{Row Total} * \text{Column Total}}{\text{Grand Total}}$)

Expected Frequencies are calculated only for the cells containing correct predictions.

Cohen's Kappa Metric

- Total= $30+357+9+2170=2566$
- Observed Accuracy**= $(30+2170)/2566=0.857$
- Expected True Positives= $(387*39)/2566 = 5.88$
- Expected True Negatives= $(2179*2527)/2566 = 2145.88$
- Expected Accuracy by random chance**= $(5.88+2145.88)/2566 = 0.839$

$$kappa = \frac{Observed\ Accuracy - Expected\ Accuracy}{1 - Expected\ Accuracy} = \frac{0.857 - 0.839}{1 - 0.839} = \frac{0.018}{0.161} = 0.11$$

Slightly better than random chance!

$$\text{Accuracy} = \frac{\text{True +ve} + \text{True-ve}}{\text{Total}}$$

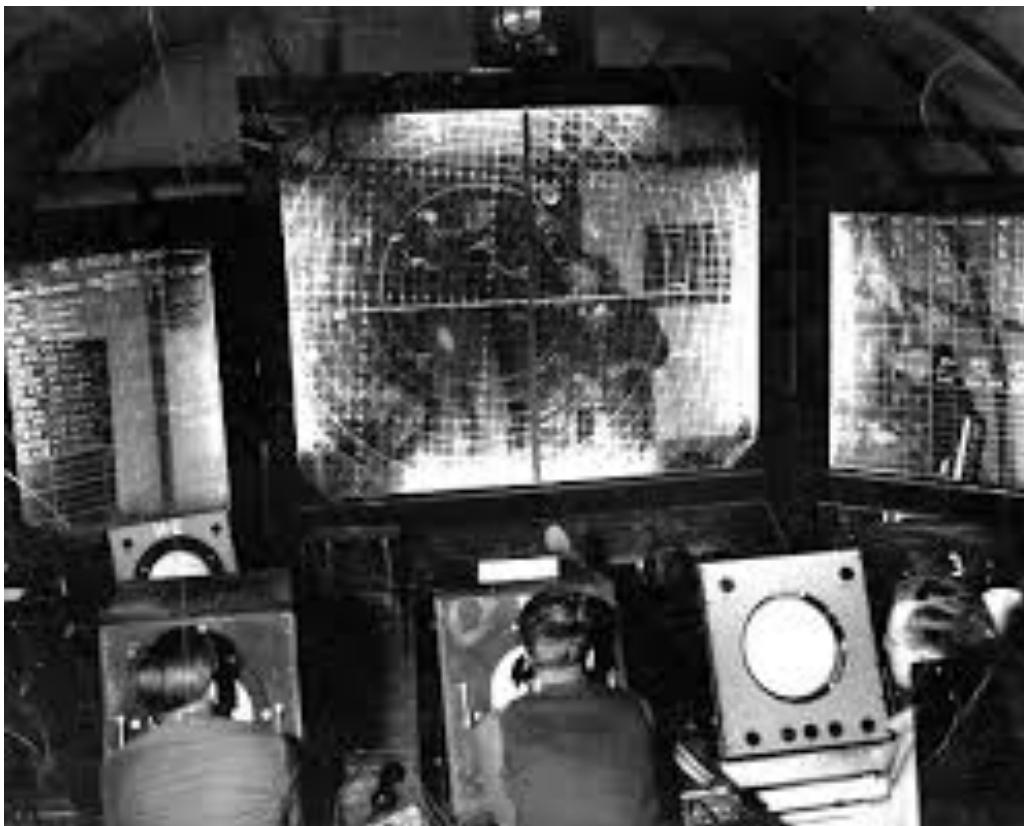
		Predicted	
		True	False
Actual	True	30	357
	False	9	2170

Kappa Value	Interpretation*
<0	No agreement
0-0.2	Slight
0.21 to 0.4	Fair
0.4 to 0.6	Moderate
0.6 to 0.8	Substantial
0.8 to 1	Almost Perfect

* Landis, J.R.; Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics* 33 (1): 159–174

ROC Curves and AUC

- ROC – Receiver Operating Characteristics
- AUC – Area Under the ROC Curve



Confusion Matrix - Metrics

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False -ve	Recall/Sensitivity/True Positive Rate (Minimize False -ve)
	Negative	False +ve	True -ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{Actual+ve}}$$

$$\text{Recall (Sensitivity)} = \frac{\text{True+ve}}{\text{True+ve} + \text{False-ve}}$$

$$\text{Specificity} = \frac{\text{True -ve}}{\text{Actual-ve}}$$

$$\text{Specificity} = \frac{\text{True -ve}}{\text{False+ve} + \text{True -ve}}$$

$$\text{Precision} = \frac{\text{True +ve}}{\text{Predicted +ve}}$$

$$\text{Precision} = \frac{\text{True+ve}}{\text{True+ve} + \text{False+ve}}$$

$$\text{Accuracy} = \frac{\text{True +ve} + \text{True-ve}}{\text{Total}}$$

$$\text{Accuracy} = \frac{\text{True+ve} + \text{True -ve}}{\text{True+ve} + \text{False-ve} + \text{False+ve} + \text{True -ve}}$$

$$F_1 \text{ Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs (1-Specificity)

Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	True Positives	False Positives	True Negatives	False Negatives
0.9	0	0	922	170
0.7	1	1	921	169
0.5	13	6	916	157
0.3	45	80	842	125
0.1	139	473	449	31

- Actual Counts

- Without CHD: 922
- With CHD: 170

> `table(test$TenYearCHD, predictTest > 0.5)`

	FALSE	TRUE
0	916	6
1	157	13

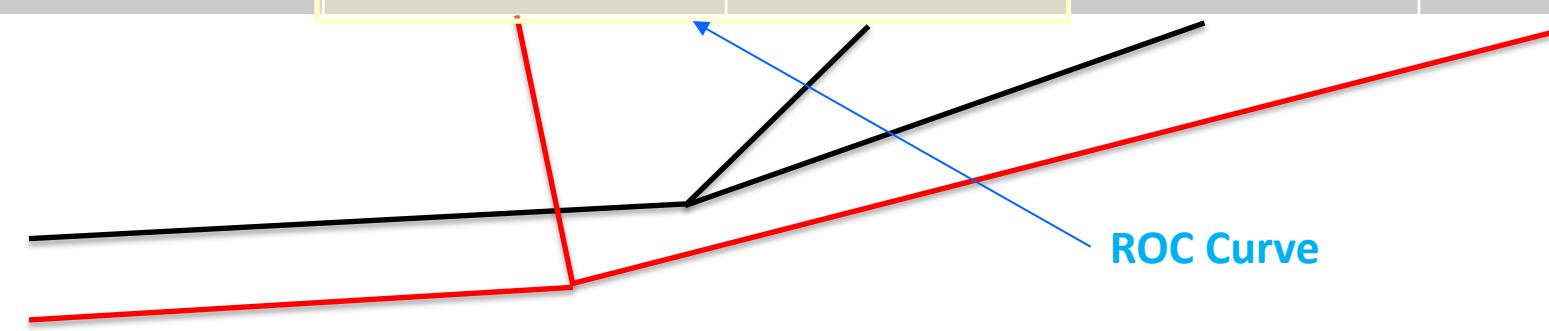


ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity

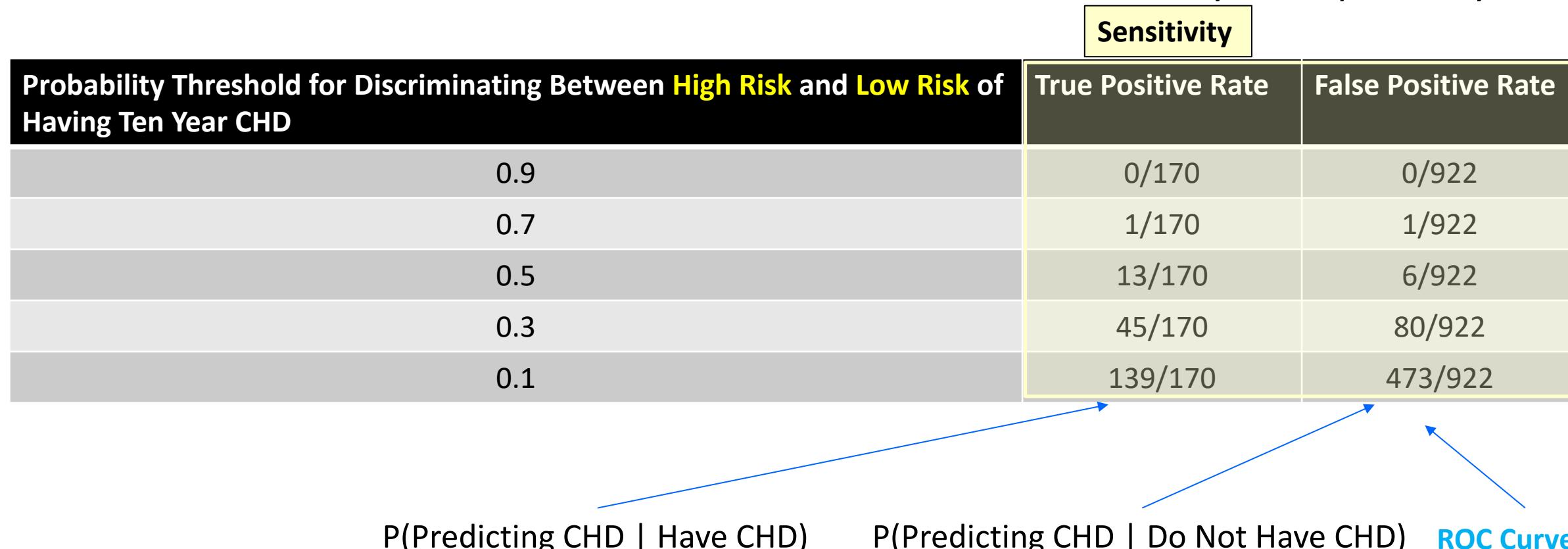
Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	Sensitivity	Specificity		
	True Positive Rate	False Positive Rate	True Negative Rate	False Negative Rate
0.9	0/170	0/922	922/922	170/170
0.7	1/170	1/922	921/922	169/170
0.5	13/170	6/922	916/922	157/170
0.3	45/170	80/922	842/922	125/170
0.1	139/170	473/922	449/922	31/170

- Actual Counts
 - Without CHD: 922
 - With CHD: 170



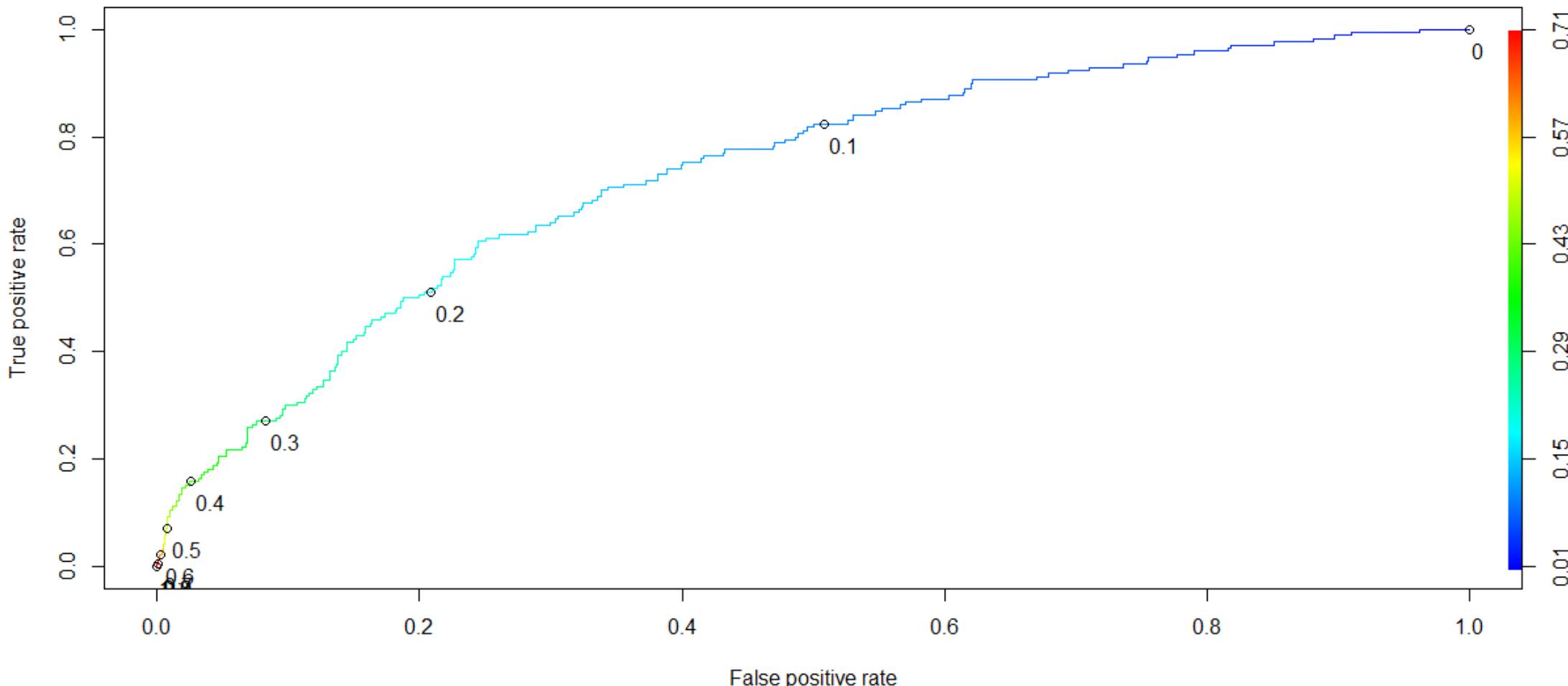
ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity



ROC Curves and AUC

- ROC – Plot of True Positive Rate vs False Positive Rate, i.e., Sensitivity vs 1-Specificity

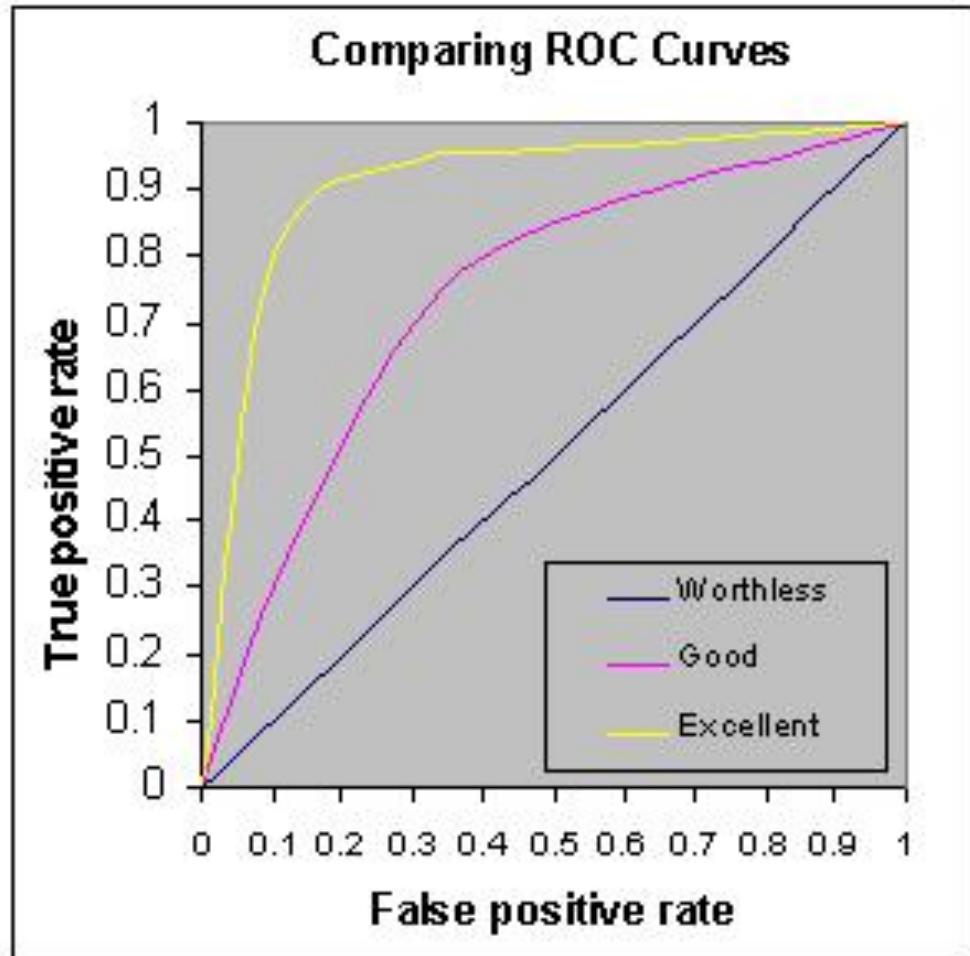


ROC Curves and AUC

- AUC – Measures discrimination, i.e., ability to correctly classify those with and without CHD.
- If you randomly pick one person who HAS CHD and one who DOESN'T and run the model, the one with the higher probability should be from the high risk group.
- AUC is the percentage of randomly drawn such pairs for which the classification is done correctly.



ROC Curves and AUC



Rough rule of thumb:

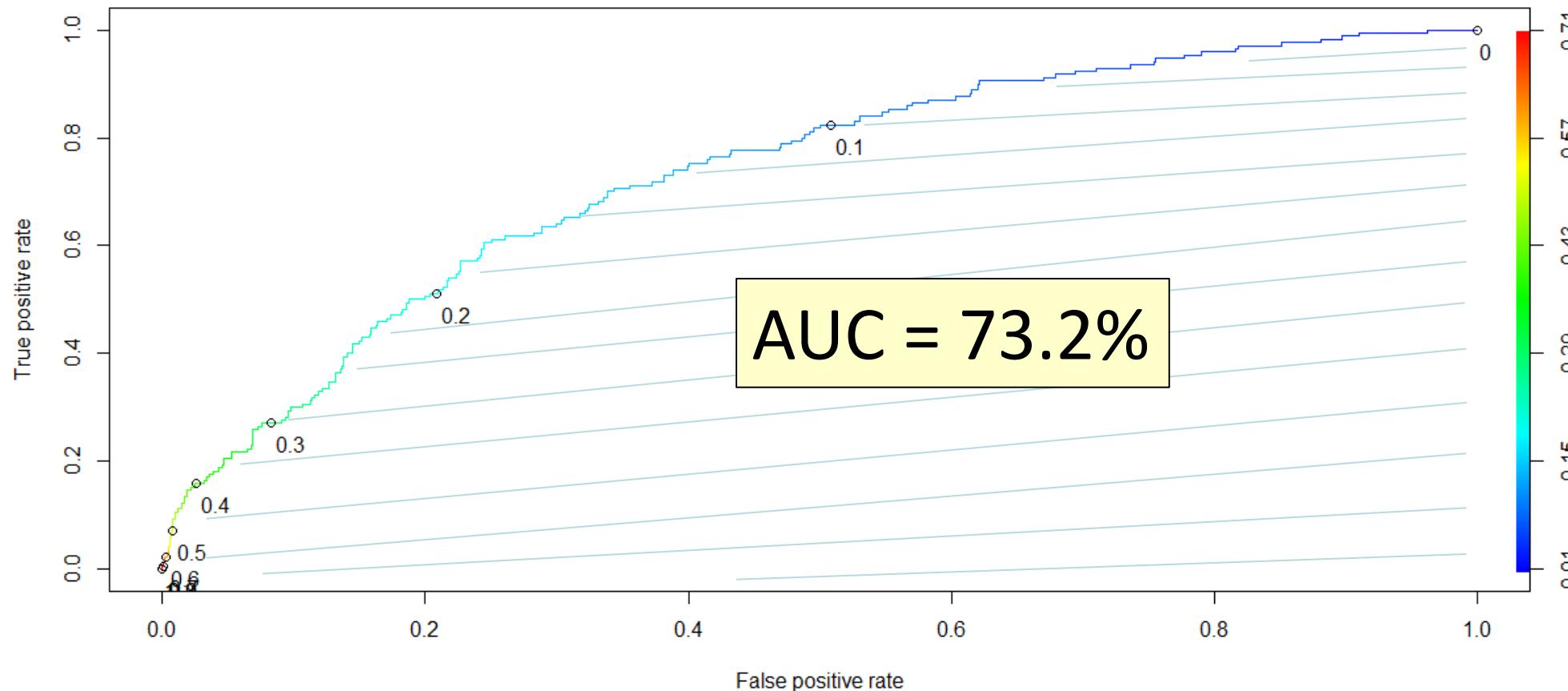
- 0.90 - 1.0 = Excellent
- 0.80 – 0.90 = Good
- 0.70 – 0.80 = Fair
- 0.60 – 0.70 = Poor
- 0.50 – 0.60 = Fail

- <0.50 – You are better off doing a coin toss than working hard to build a model 😊

ROC Curves and AUC

BREAK

- The model does a fair job of discrimination between high risk and low risk people.
- Useful for comparing different models.





Classification

NAÏVE BAYES ALGORITHM

Classification Problems with Multiple Classes

- Given an article, predict which section of the newspaper (Current News, International, Arts, Sports, Fashion, etc.) it is supposed to go to
- Given a photo of a car number plate, identify which state it belongs to
- Given an audio clip of a song, identify the genre
- Given an email, predict whether it is spam or not spam (a 2-class problem)



Classification Problems

- All classification problems are essentially equivalent to evaluating conditional probability
- $P(Y_i | X)$, i.e., given certain evidence X , what is the probability that this is from class Y_i
- Logistic Regression solves this problem by modelling the probabilistic relationship between X and Y (sigmoid function, linear in X , etc.) **directly**
- Such models are called Discriminative Models



Naïve Bayes Algorithm

- Naïve Bayes makes predictions - $P(Y_i | X)$ - using Bayes theorem after modelling the joint probability of X and Y

$$\text{Recall Conditional Probability} = \frac{\text{Joint Probability}}{\text{Marginal Probability}}$$

- These type of methods are called Generative Learning Models
- A simple classifier that performs surprisingly well on a large class of problems



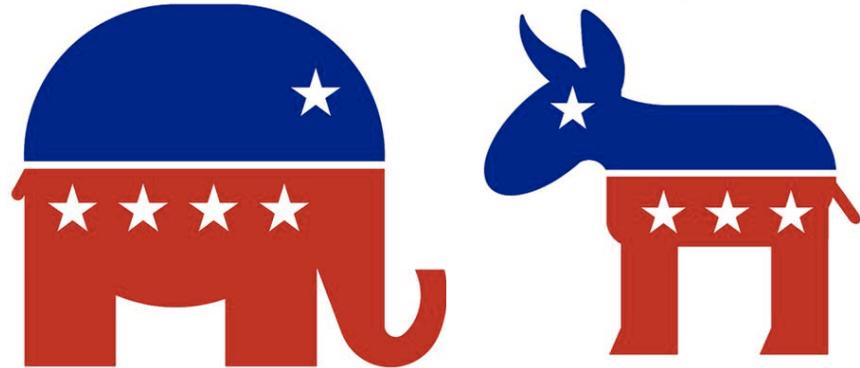
US House of Congress Voting Patterns

Class	v1	v2	v3	v4	v5	v6	v7	1	Class Name: 2 (democrat, republican)
republican	n	y	n	y	y	y	n	2	handicapped-infants: 2 (y,n)
republican	n	y	n	y	y	y	n	3	water-project-cost-sharing: 2 (y,n)
democrat	NA	y	y	NA	y	y	n	4	adoption-of-the-budget-resolution: 2 (y,n)
democrat	n	y	y	n	NA	y	n	5	physician-fee-freeze: 2 (y,n)
democrat	y	y	y	n	y	y	n	6	el-salvador-aid: 2 (y,n)
democrat	n	y	y	n	y	y	n	7	religious-groups-in-schools: 2 (y,n)
republican	n	y	n	y	y	y	n	8	anti-satellite-test-ban: 2 (y,n)
republican	n	y	n	y	y	y	n	9	aid-to-nicaraguan-contras: 2 (y,n)
democrat	y	y	y	n	n	n	y	10	mx-missile: 2 (y,n)
republican	n	y	n	y	y	n	n	11	immigration: 2 (y,n)
republican	n	y	n	y	y	y	n	12	synfuels-corporation-cutback: 2 (y,n)
democrat	n	y	y	n	n	n	y	13	education-spending: 2 (y,n)
democrat	y	y	y	n	n	y	y	14	superfund-right-to-sue: 2 (y,n)
republican	n	y	n	y	y	y	n	15	crime: 2 (y,n)
democrat	y	n	y	n	n	y	n	16	duty-free-exports: 2 (y,n)
republican	n	y	n	y	y	y	n	17	export-administration-act-south-africa: 2 (y,n)
democrat	y	y	y	n	n	n	y		

House Votes 1984 Dataset: Voting patterns of Members of Congress.

A data frame with 435 observations on 17 variables. 168 Republicans, 267 Democrats

Republican or Democrat?



Republican – R – Red

Democrat – D - Donkey

Given a Congressman's voting pattern ($v1 = y$, $v2 = n$), what is the probability that this person is a Democrat?

$$P(D \mid v1 = y, v2 = n) = ?$$

Prior Belief - Simplest Solution

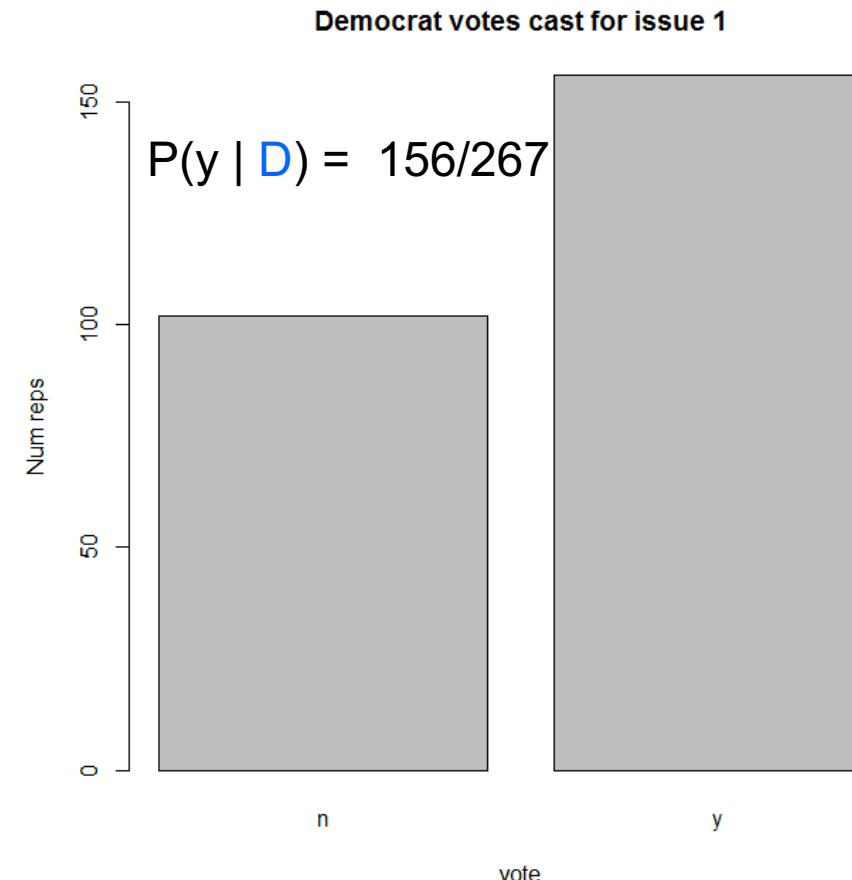
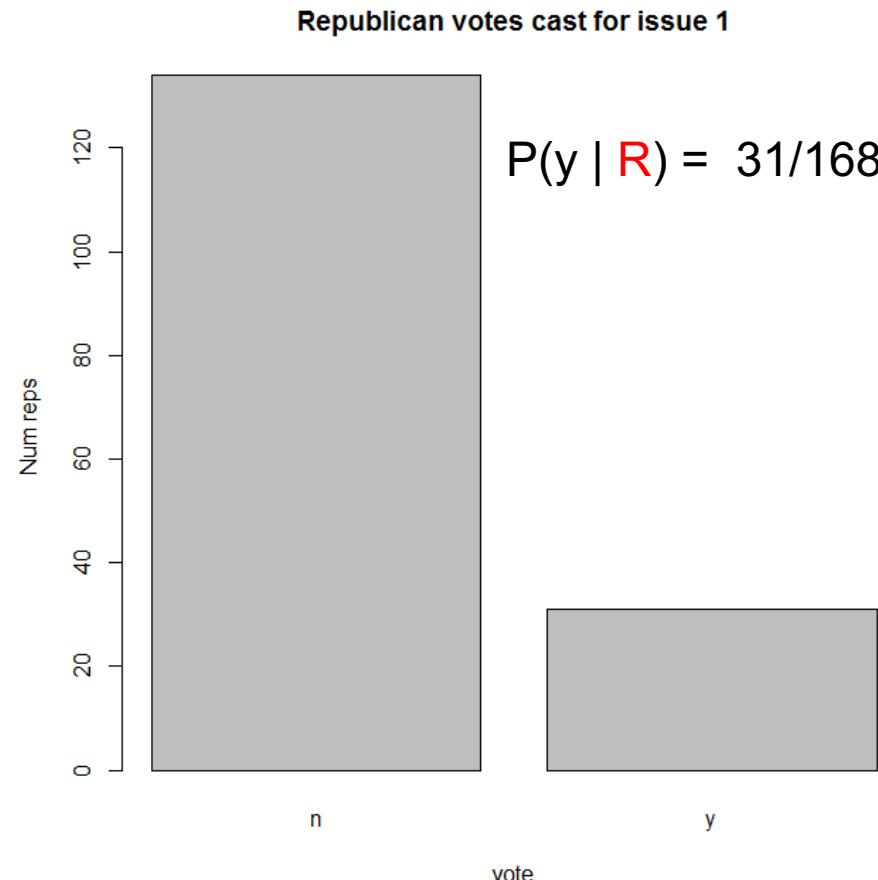
- The house has a majority of Democrats
 - 168 Republicans, 267 Democrats
- Probability of a random person being Democrat is
 - $P(D) = 267/435 = 0.61$
- Can we do better by incorporating the evidence of their voting patterns?

Voting Patterns for V1

Handicapped Infants. The vote failed to pass: 236 to 187

We are interested in $P(D | v1 = y, v2 = n) = ?$

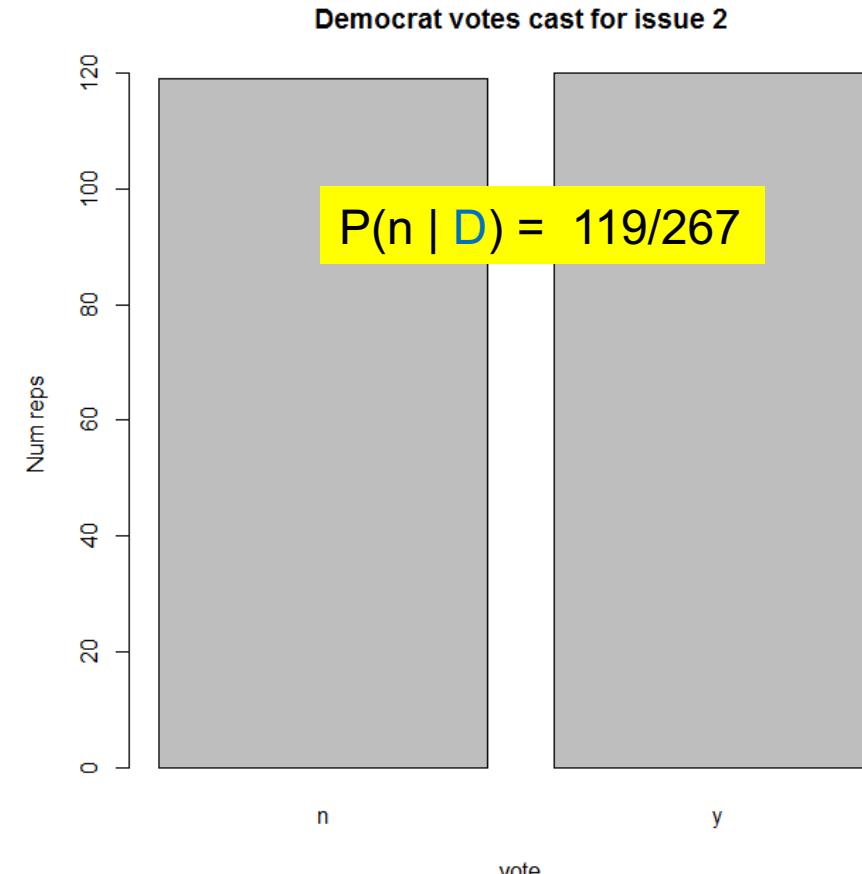
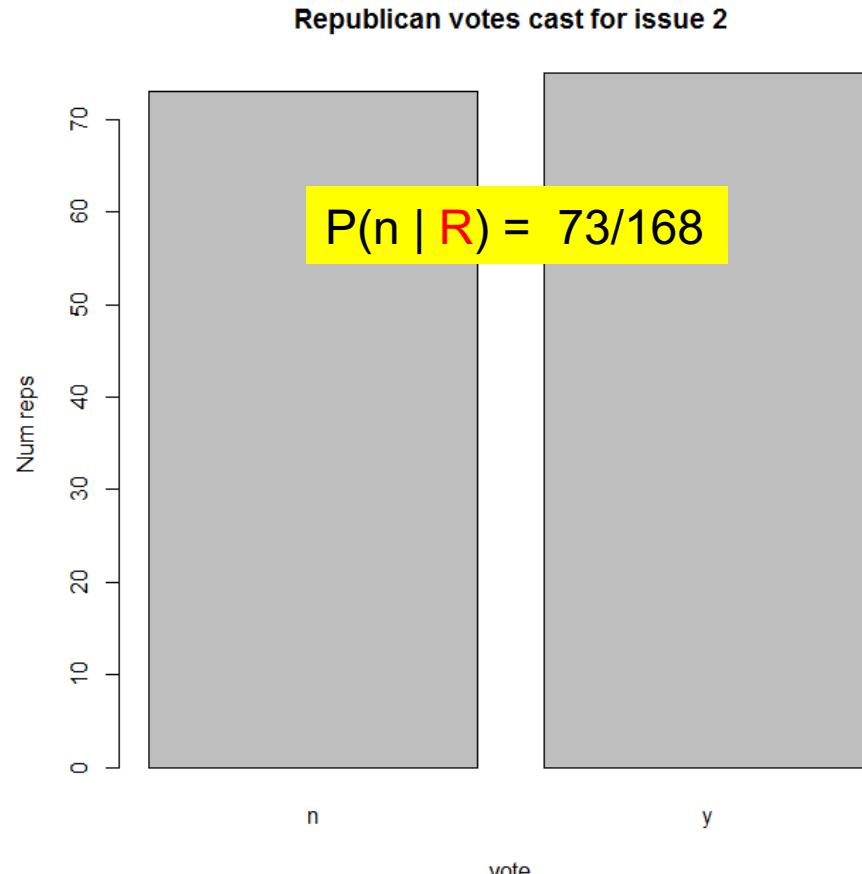
```
> Repub <- HouseVotes84$Class=="republican"  
> Democrat <- HouseVotes84$Class=="democrat"  
> plot(as.factor(HouseVotes84[Repub,2]))  
> title(main="Republican votes cast for issue 1",  
xlab="vote", ylab="Num reps")  
> plot(as.factor(HouseVotes84[Democrat,2]))  
> title(main="Democrat votes cast for issue 1", x  
lab="vote", ylab="Num reps")
```



Voting Patterns for V2

Water-project-cost-sharing. The vote passed: 195 to 192

We are interested in
 $P(D | v1 = y, \text{v2} = n) = ?$



Bayes Theorem

$$P(A|B) = ?$$

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

$$P(D|v1 = y, v2 = n) = ?$$

$$P(D|v1 = y, v2 = n) = \frac{P(D) * P(v1 = y, v2 = n|D)}{P(v1 = y, v2 = n)}$$



Naïve Bayes

Naïve Assumption: *Conditional probability of each feature given the class, is independent of all other features*

$$P(v1 = y, v2=n | D) = P(v1 = y|D) * P(v2 = n|D)$$

$$P(D|v1 = y, v2=n) = \frac{P(D) * P(v1 = y|D) * P(v2 = n|D)}{P(v1 = y, v2=n)}$$

Naïve Bayes

We are trying to decide, given the voting pattern, if that person is a Democrat or a Republican.

$$P(D|v1 = y, v2=n) = \frac{P(D) * P(v1 = y|D) * P(v2 = n|D)}{P(v1 = y, v2=n)}$$

$$P(R|v1 = y, v2=n) = \frac{P(R) * P(v1 = y|R) * P(v2 = n|R)}{P(v1 = y, v2=n)}$$

Whichever probability is higher, we would classify the person into that party.

Note that the denominator is the same for both. So we need to focus only on numerator.

Naïve Bayes

$$P(D|v1 = y, v2=n) \propto P(D) * P(v1 = y|D) * P(v2 = n|D)$$

$P(D) = 267/435$ (267 Democrats among 435 Congressmen)

$$P(D|v1 = y, v2=n) \propto \frac{267}{435} * \frac{156}{267} * \frac{119}{267} = 0.15$$

From voting pattern
[slide](#)

$$P(R|v1 = y, v2=n) \propto \frac{168}{435} * \frac{31}{168} * \frac{73}{168} = 0.03$$

Since the conditional probability for being Democrat is higher, he is likely to be Democrat.

Naïve Bayes: Voting patterns



```
library(e1071)
nb_model <- naiveBayes(Class~., data = trainHouseVotes84)

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
democrat republican
0.6111111 0.3888889

Conditional probabilities:
V1
Y          n      y
democrat 0.4066986 0.5933014
republican 0.8195489 0.1804511

V2
Y          n      y
democrat 0.5119617 0.4880383
republican 0.4586466 0.5413534
```

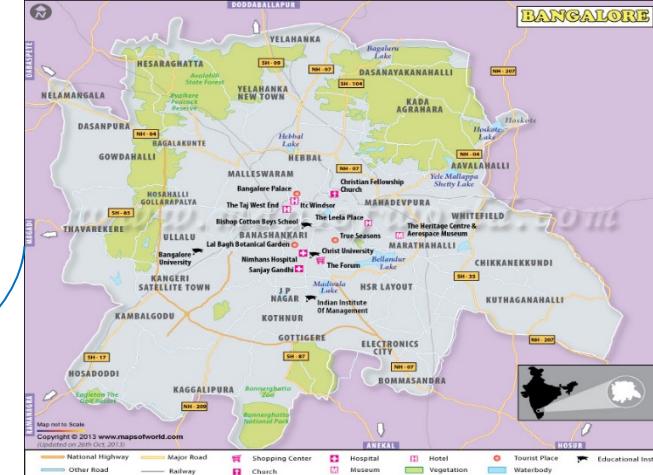
Naïve Bayes Assumption

- The key assumption of independence of features, is almost never true
- Still Naïve Bayes does surprisingly well in a lot of situations
- It works best when all the predictor variables are categorical variables
- Very frequently used in text mining, character image analysis problems





Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
 +91-9701685511 (Individuals)
 +91-9618483483 (Corporates)

BENGALURU

Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
 +91-9502334561 (Individuals)
 +91-9502799088 (Corporates)

Social Media

- Web: <http://www.insofe.edu.in>
- Facebook: <https://www.facebook.com/insofe>
- Twitter: <https://twitter.com/Insofeedu>
- YouTube: <http://www.youtube.com/InsofeVideos>
- SlideShare: <http://www.slideshare.net/INSOFE>
- LinkedIn: <http://www.linkedin.com/company/international-school-of-engineering>

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.