Inspire...Educate...Transform.






Inspire…Educate…Transform.

Statistics and Probability in Decision Modeling

Logistic Regression and Naïve Bayes

Dr. Anand Narasimhamurthy

Acknowledgements : A number of slides are due to Dr.Sridhar Pappu

Outline

- Motivation for and basics of logistic regression with examples
- Understanding the nuts and bolts of logistic regression
  - Background : Link function, log odds
  - Reading and interpreting model output
- Classification performance evaluation concepts
  - Confusion matrix, Sensitivity, Specificity, ROC
  - Gain and lift charts
- Naive Bayes
  - Review of Bayes Theorem
  - Understanding the “Naive” part of Naive Bayes

Logistic regression : Overview

- A statistical technique developed by David Cox in 1958.
- Logistic regression is **usually** taken to apply to a **binary dependent variable**, though it can be extended to variables with multiple categories.
- Inputs are one or more predictor variables
- The model itself **outputs a conditional probabilit**y of output in terms of input, however it is usually straightforward to convert this to a classification output.

Examples of two class classification problems

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

Examples of classification problems with multiple classes

- Given an article – predict which section of the newspaper (Current News, International, Arts, Sports,Fashion etc) it supposed to go
- Given a photo of a car number plate, identify which state it belongs to
- Audio clip of a song, identify the genre

Background concepts and notation

- Consider a two class classification problem. Let the class labels be coded as 1 and 0.
  - Let 1 denote the target class.
- Let p denote the probability of occurence of target class, given predictor variables.
  i.e. p = conditional probability $P(Y=1 \mid x)$ where x could represent one or more predictors.
- The **odds of success** (sometimes just called odds) is defined as :

$$\textbf{odds} = \mathbf{p/(1-p)}$$

- Sometimes it is convenient to use log-odds which is defined as the natural log (log to base e) of the odds. This is often referred to as the logit function :

$$\textbf{logit}(\mathbf{p}) = \mathbf{ln\ (p/(1-p))}$$

CSE 7202c

Probability and Odds : Examples

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 19:2, what is the probability of winning?	19/21
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

Illustrative examples where logistic regression may be suitable

CSE 7202c

Example 1 : Predict approval based on credit score

creditScore	approved
655	0
692	0
681	0
663	1
688	1
693	1
699	0
699	1
683	1
698	0
655	1
703	0
704	1
745	1
702	1

$n = 1000$

creditScore is the applicant's credit score

approved is coded "1" for approved and "0" for not approved; it is a binary, mutually exclusive variable.

* Only 15 of 1000 observations shown

Source : Brandon Foltz, Logistic Regression youtube

Aim : Develop a model that takes a given credit score as input and computes the probability (and hence odds) of being approved.

Typical questions one would like to answer using the model

Discover from the data approximately what credit score is associated with a probability of 50% of being approved (odds are even).

Predict from the model how improving credit score from say 720 to say 750 affects the probability of being approved.

Source : Brandon Foltz,Logistic Regression youtube

Example 2 : Auto club mailer flier

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?
- Additionally, can the model compute the probability (and hence odds) of a particular member returning the form?

Automailer : Snapshot of data

Age	Response
...	...
50	1
51	1
64	1
54	1
52	0
42	0
45	0
33	0
...	...

Response categories coded as below.

Yes : 1
No : 0

Example 3 – 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.
- Study led to development of **Framingham Risk Score**, a gender specific algorithm used to **estimate the 10-year cardiovascular risk of an individual**: https://www.framinghamheartstudy.org/ http://cvdrisk.nhlbi.nih.gov/

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
  - *prevalentStroke*: Did the subject have a previous stroke – Yes or No
  - *prevalentHyp*: 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)

Particularly useful model outputs would be :

- A risk score in addition to a classification of risk category (Low/Medium/High)

- Significant variables that can be controlled eg.
  - Smoking habits
  - Cholesterol
  - Systolic BP
  - Blood glucose

  and the impact each has on the odds of CHD

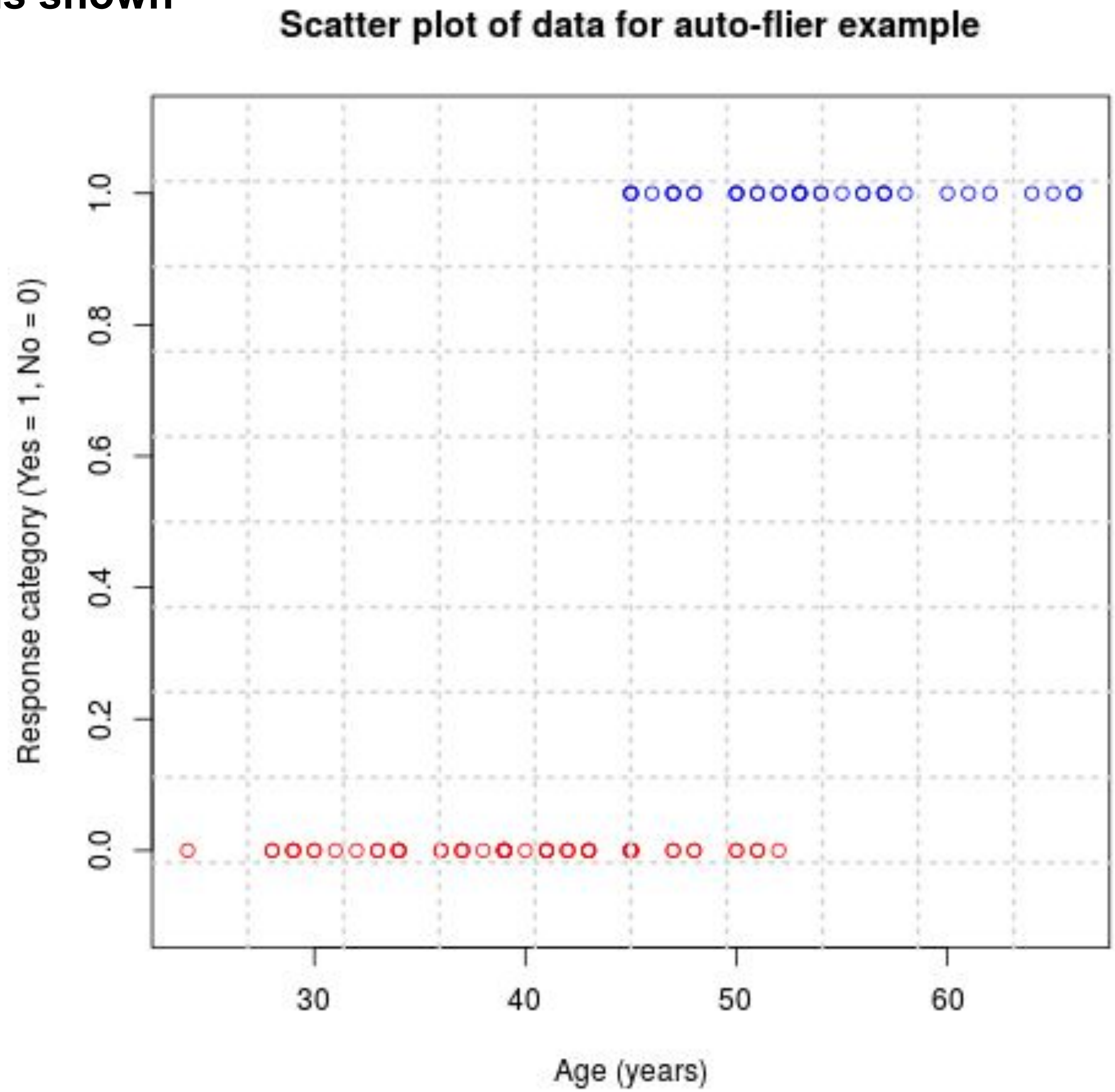
Building a logistic regression model

Example 2 : Auto club mailer flier

CSE 7202c

Table of data, only few records shown

Age	Response
...	...
50	1
51	1
64	1
54	1
52	0
42	0
45	0
33	0
...	...



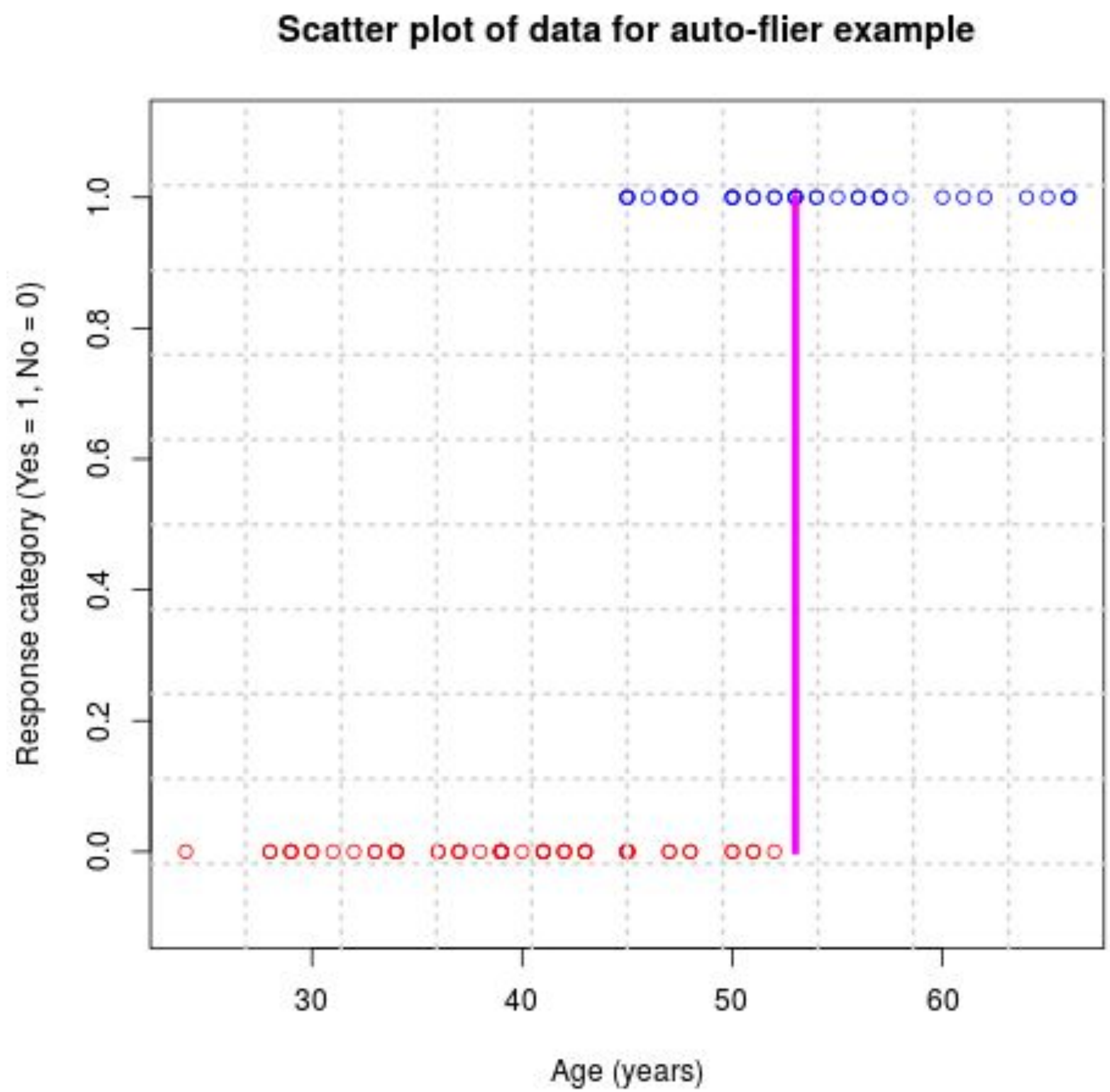
Building a model for the auto mailer example using a single predictor (age)

Problem type : Classification
Inputs : One numeric variable (age)
Output : Response category (with two levels)

Hence the problem is a Binary (two-class) classification problem.

Proposed models :
Model 1 : A 1D classifier, i.e. a threshold (Lab)
Model 2 : Logistic regression (Class and lab)

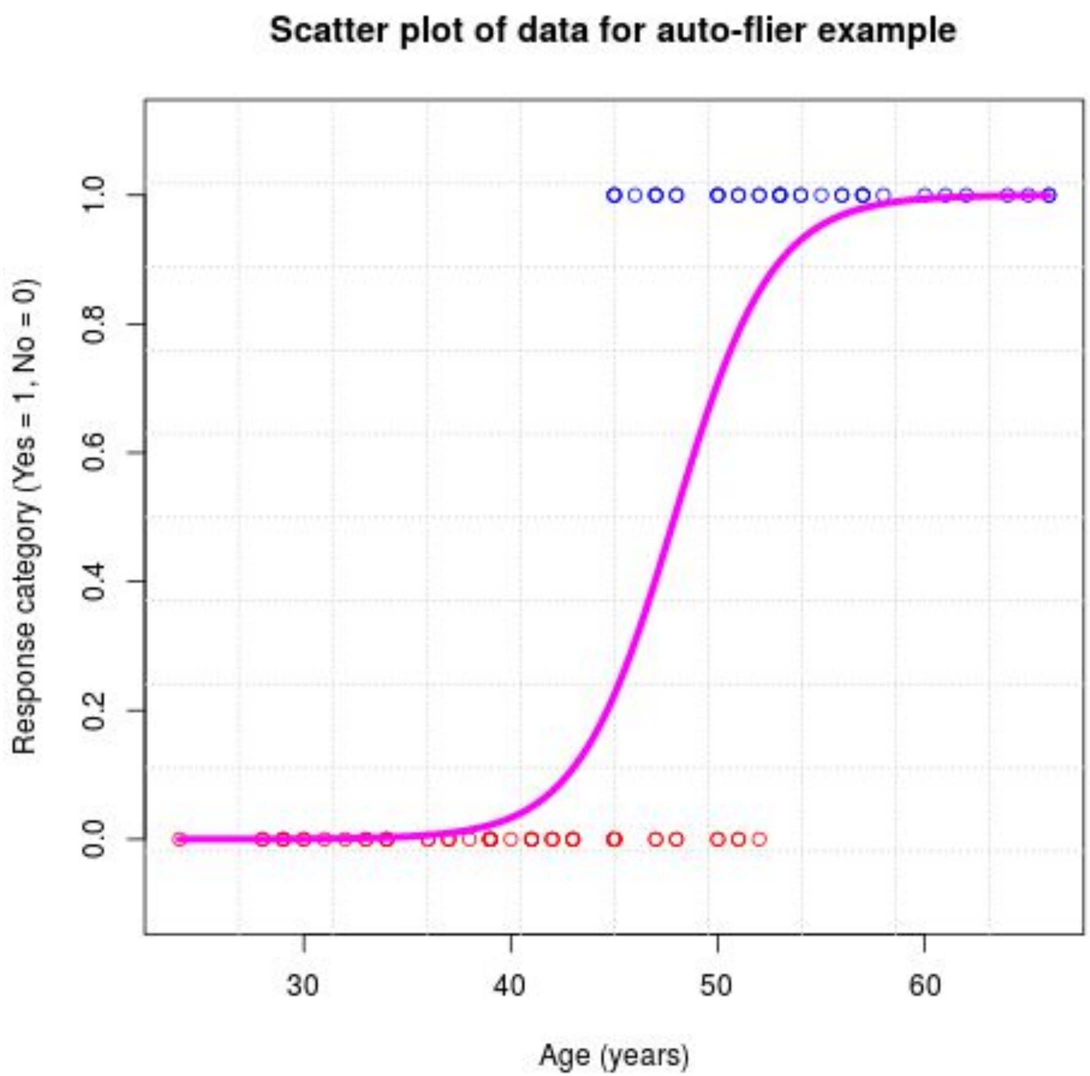
Model 1 : A 1D classifier (threshold)



Model parameters :
A single threshold (shown as a magenta line)

Classification rule :
If Age > threshold, classify as 1
(i.e. Predict that person will respond)
otherwise, classify as 0

Model 2 : A logistic regression model

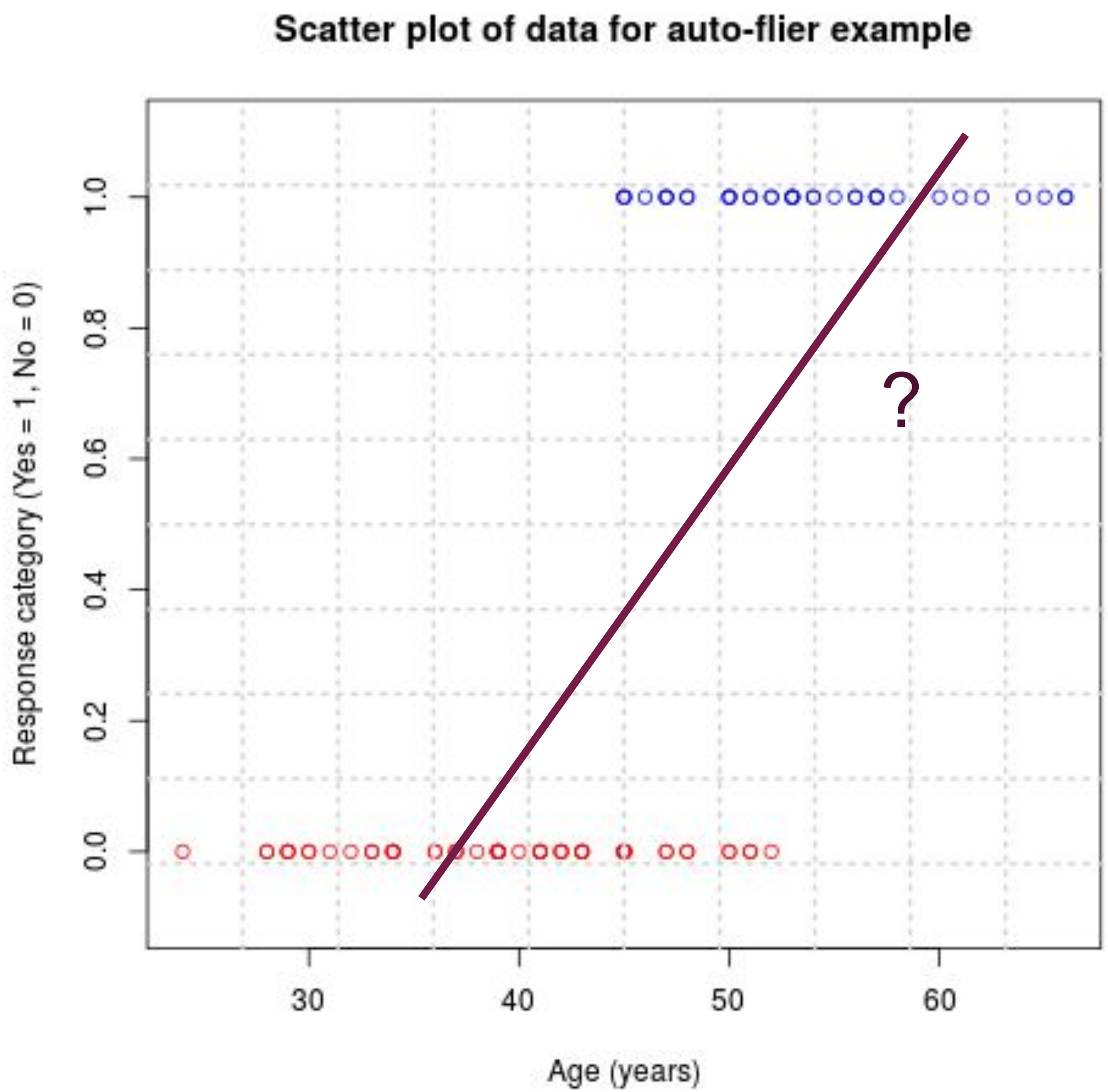


Model inputs :
Value of one or more predictors

Model output :
Prob(Class = Target | predictor values) = P(Y=1 | $x_1,x_2,...,x_k$)

(In our example, Age is the only predictor).

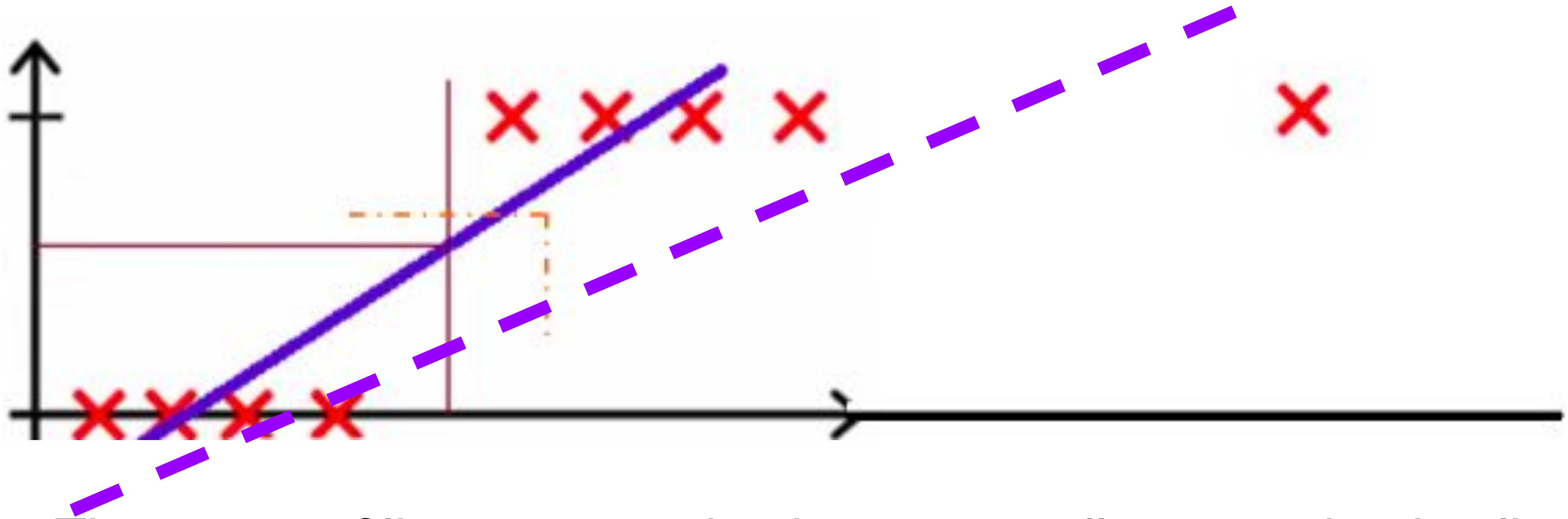
Classification Tasks: Can regression be used?



Consider a two class problem, where the class labels are coded as 0 and 1

How about using linear regression for classification with target values as the class labels 0 and 1?

Linear regression for classification has major drawbacks



- The output of linear regression is not naturally constrained to lie within a range.

- Linear regression slopes can vary significantly depending on the data and hence thresholding becomes difficult.

Other reasons why Linear regression is not suitable

Basic assumptions of linear regression are clearly violated.

- The target variable clearly does not follow a normal distribution (binomial in our case).
- Error terms do not follow normal distribution.
- Error variances are heteroscedastic.

Hence, Linear Regression via Least Squares is inappropriate.

Logistic model

Predictor variables

$x_1, x_2, \ldots, x_k$ $\xrightarrow{\text{Linear combination}}$ $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ $\xrightarrow{\text{Sigmoid function}}$ $E(Y \mid x_1, x_2, \ldots, x_k) = P(Y=1 \mid x_1, x_2, \ldots, x_k) =$ p (say)

Sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}} = p$$

$$z = \sigma^{-1}(p) = ln\left(\frac{p}{1-p}\right)$$

logit function

where,

$z = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

and

p $= P(Y=1 \mid x_1, x_2, \ldots, x_k)$

$= E(Y \mid x_1, x_2, \ldots, x_k)$

CSE 7202c

Logistic model

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

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

Logistic model can be transformed into an odds ratio:

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

Logistic model

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

$$S = \frac{\dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}{1 - \dfrac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}}$$

The log of the odds (S) is called the **logit** and the transformed model is linear in **β**s

$$\therefore, S = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$$

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

Example 2 (Flier mailer) : Model output from R

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

Individual regression coefficients

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?

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.

```
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
```

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).

Interpreting Output – Testing the Overall Model

```
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
```

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

R outputs AIC (Akaike’s Information Criterion) and you need to pick the model with the lowest AIC.

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

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

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

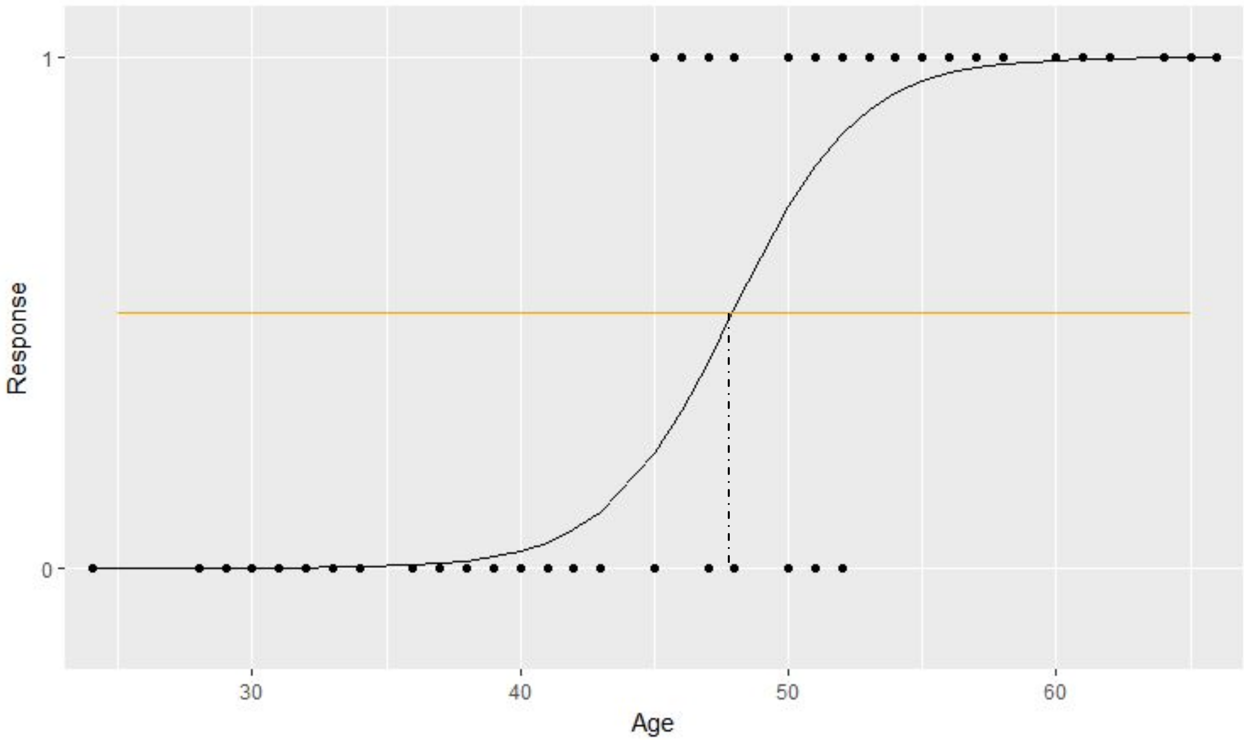
The output of the logistic regression forecast is a probability value. One needs to decide on a threshold value before a class is assigned.

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
```

Visualizing the fit



The threshold of p=0.5, corresponds to the point where Ln(S) = 0.

We can obtain the age at which
the model switches from class 0 to class 1, by setting Ln(S) to be zero in the logistic equation.

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

Setting ln(S) = 0 , we get the Age at which probability = 0.5

$Age_c = 20.40782/0.42592 = 47.9$

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.
- AIC is *similar to Adjusted* R^2 in the sense it penalizes for adding more parameters to the model.
- It offers a relative estimate of the information lost when a model is used to represent the process that generated the data.
- It does not test a model in the sense of null hypothesis and hence doesn't tell anything about the quality of the model. It is only a relative measure between multiple models.
- AIC = n Log(SSE/n) + 2k   for Ordinary Least Squares

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. $GVIF = VIF^{\left(\frac{1}{2*df}\right)}$

CSE 7202c

Performance Measures for Regression and Classification Models

CSE 7202c

A short review of performance metrics for classification

Recall concepts covered in previous sessions

- Accuracy alone can be misleading, especially if there is a class imbalance.

- Hence, depending on the application better to report additional metrics Eg. For a two class problem with a target class, report either
  - Sensitivity and specificity or
  - Precision and Recall

Note : Recall is same as sensitivity but Precision is not same as specificity

- Most of the above measures can be derived from the **Confusion Matrix**.

Kappa Metric

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

Kappa Metric

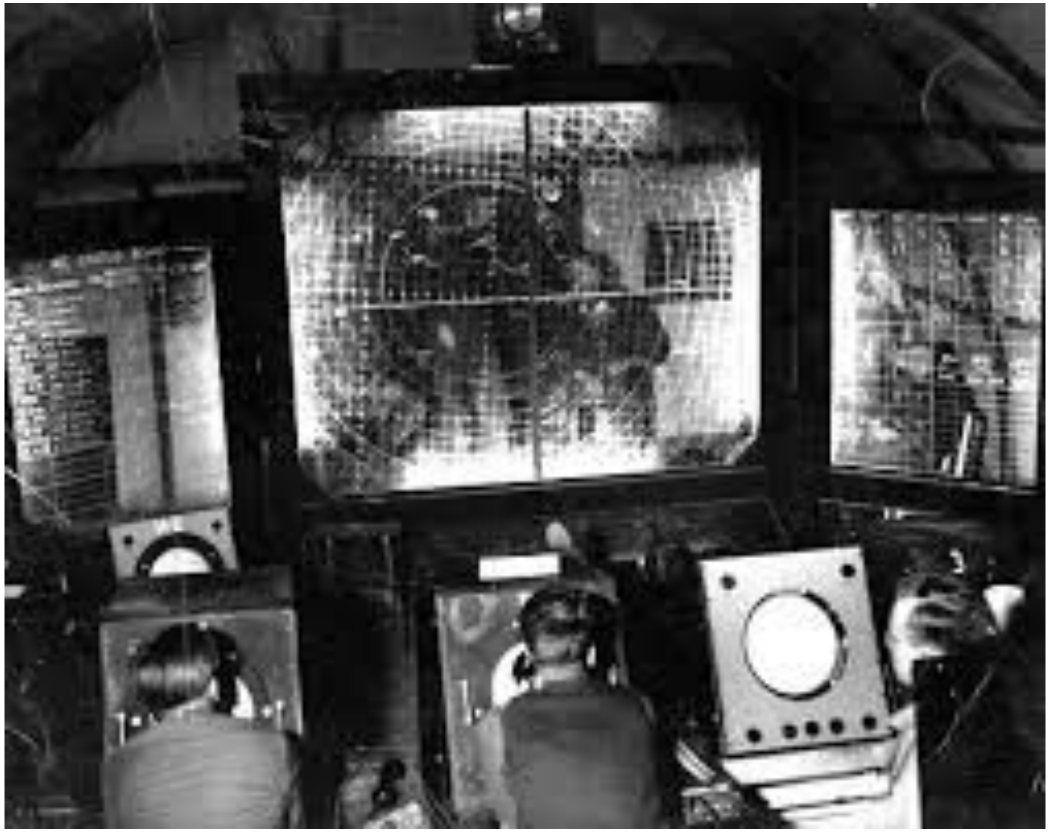
- Kappa metric quantifies how accurate the prediction algorithm is when compared to a random prediction

$$kappa = \frac{totalAccuracy - randomAccuracy}{1 - randomAccuracy}$$

Kappa Value	
<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

ROC Curves and AUC

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



- We can evaluate the classification accuracy (accuracy, sensitivity, recall, kappa etc) for a particular model (eg. a classifier at a given threshold)

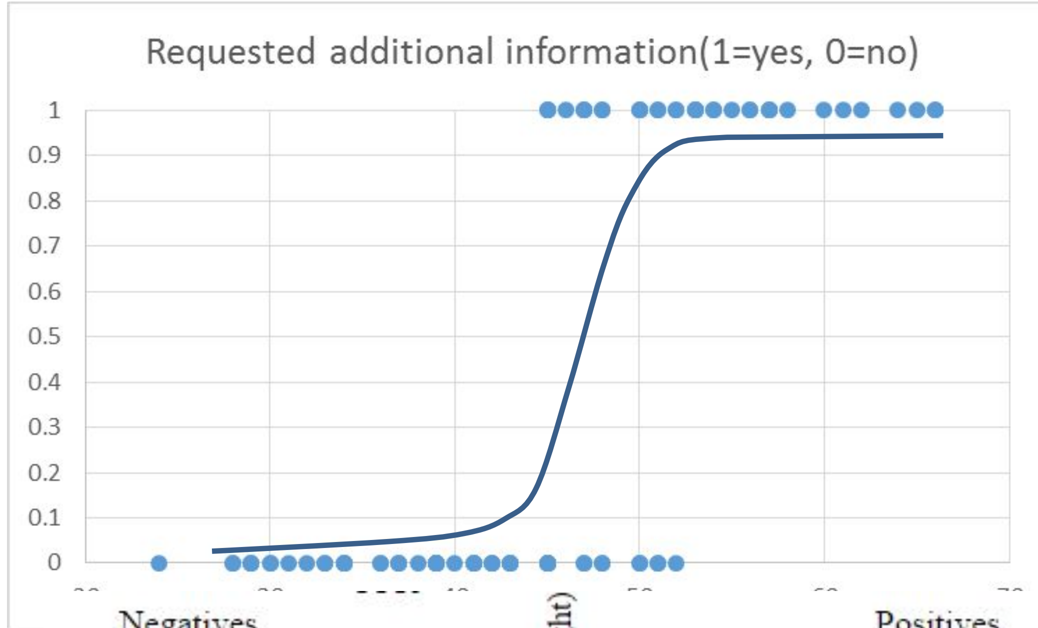
- ROC curve tries to evaluate model performance for different parameter settings (eg. at all threshold values)

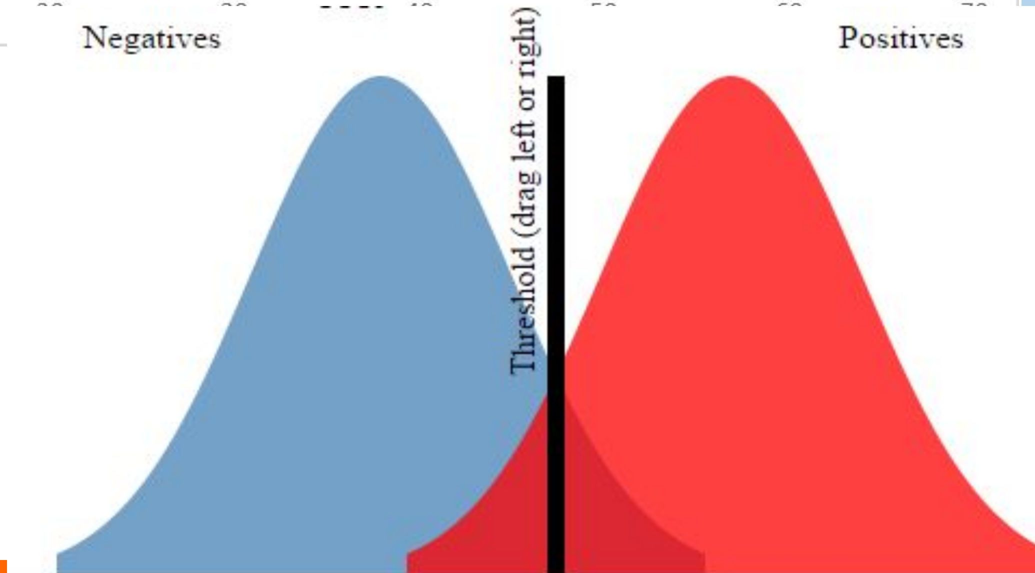
ROC Curve Demo

- http://www.navan.name/roc/
- See: https://youtu.be/OAl6eAyP-yo

Logistic regression gives Probability forecasts for the given data point to be in a given bucket.

-
- A threshold needs to be chosen to finally translate this probability to a bucket allocation





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	12	7	915	158
0.3	46	76	846	124
0.1	140	468	454	30

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 Positive Rate (Sensitivity)	False Positive Rate	True Negative Rate (Specificity)	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	12/170	7/922	915/922	158/170
0.3	46/170	76/922	846/922	124/170
0.1	140/170	468/922	454/922	30/170

ROC Curve

ROC Curves and AUC

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

Sensitivity

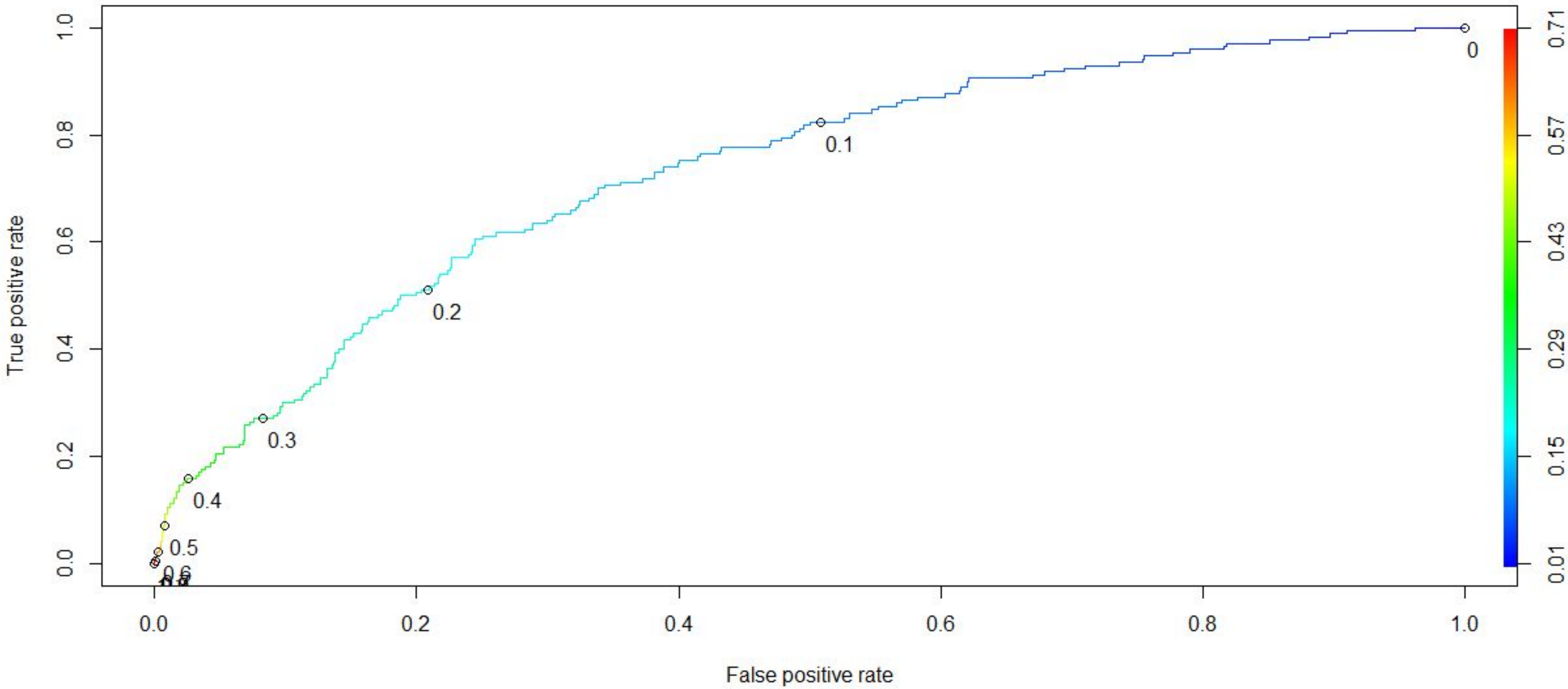
Probability Threshold for Discriminating Between High Risk and Low Risk of Having Ten Year CHD	True Positive Rate	False Positive Rate
0.9	0/170	0/922
0.7	1/170	1/922
0.5	12/170	7/922
0.3	46/170	76/922
0.1	140/170	468/922

ROC Curve

P(Predicting CHD | Have CHD) P(Predicting CHD | Do Not Have CHD)

ROC Curves and AUC

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



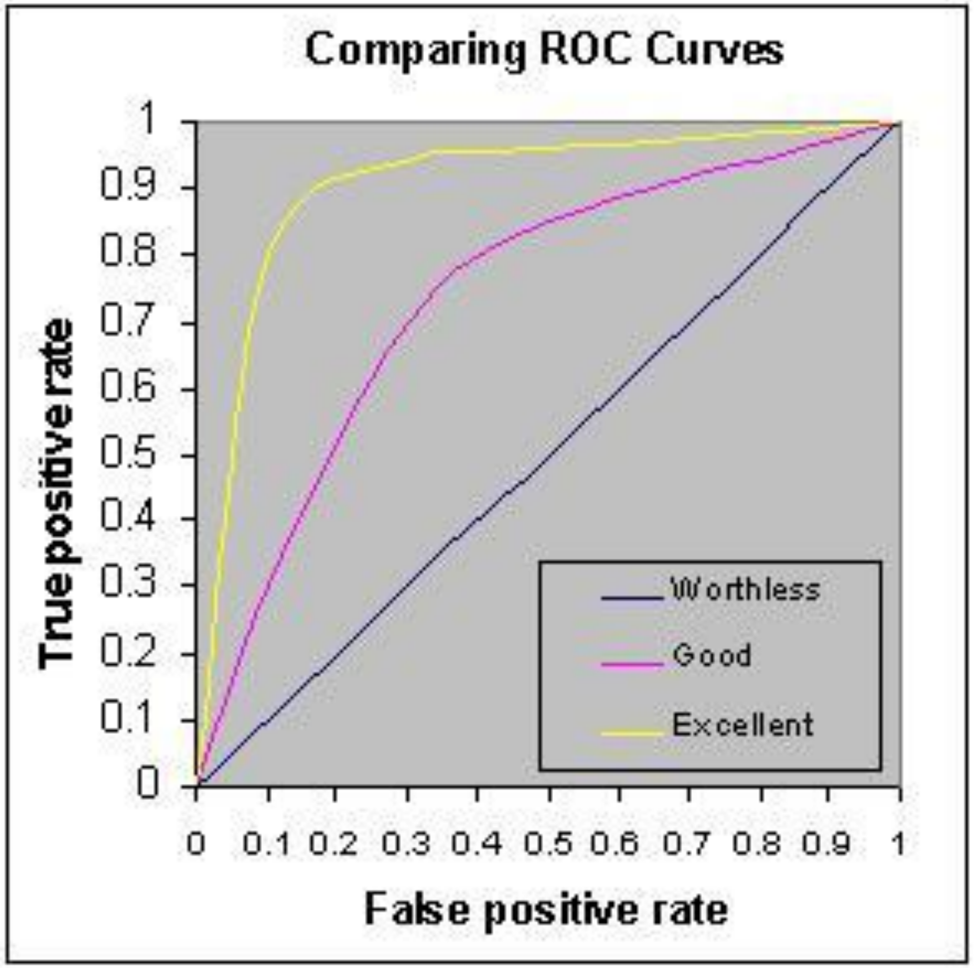
ROC Curves and AUC

- 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



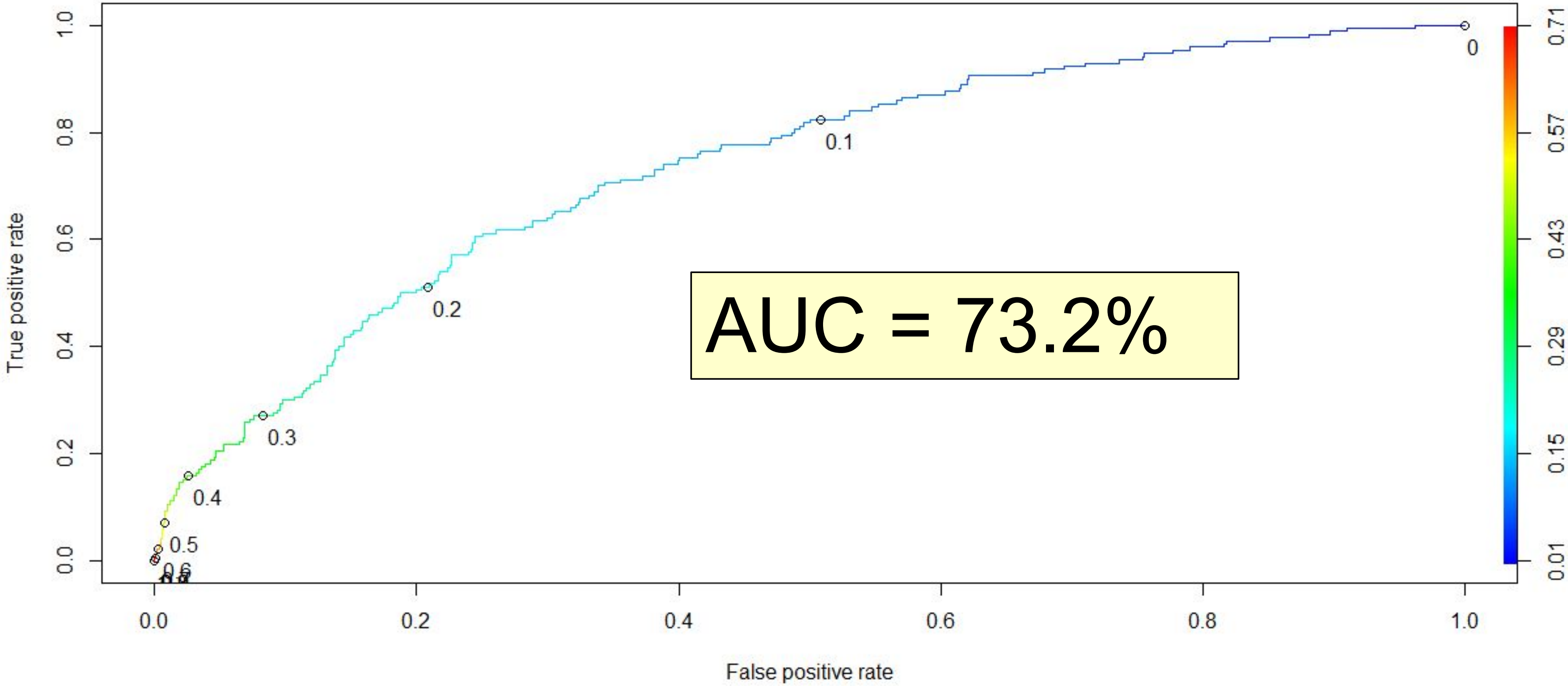
Rough rule of thumb:

- 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

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

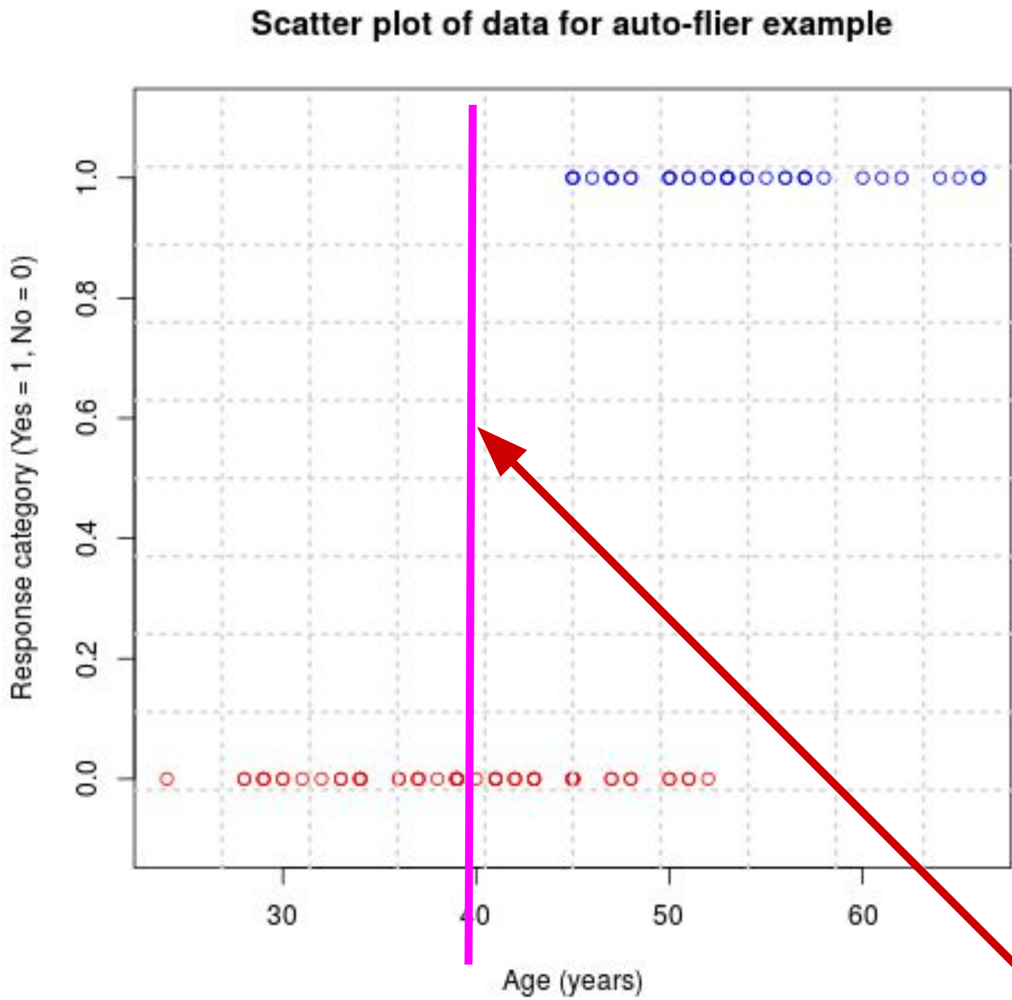


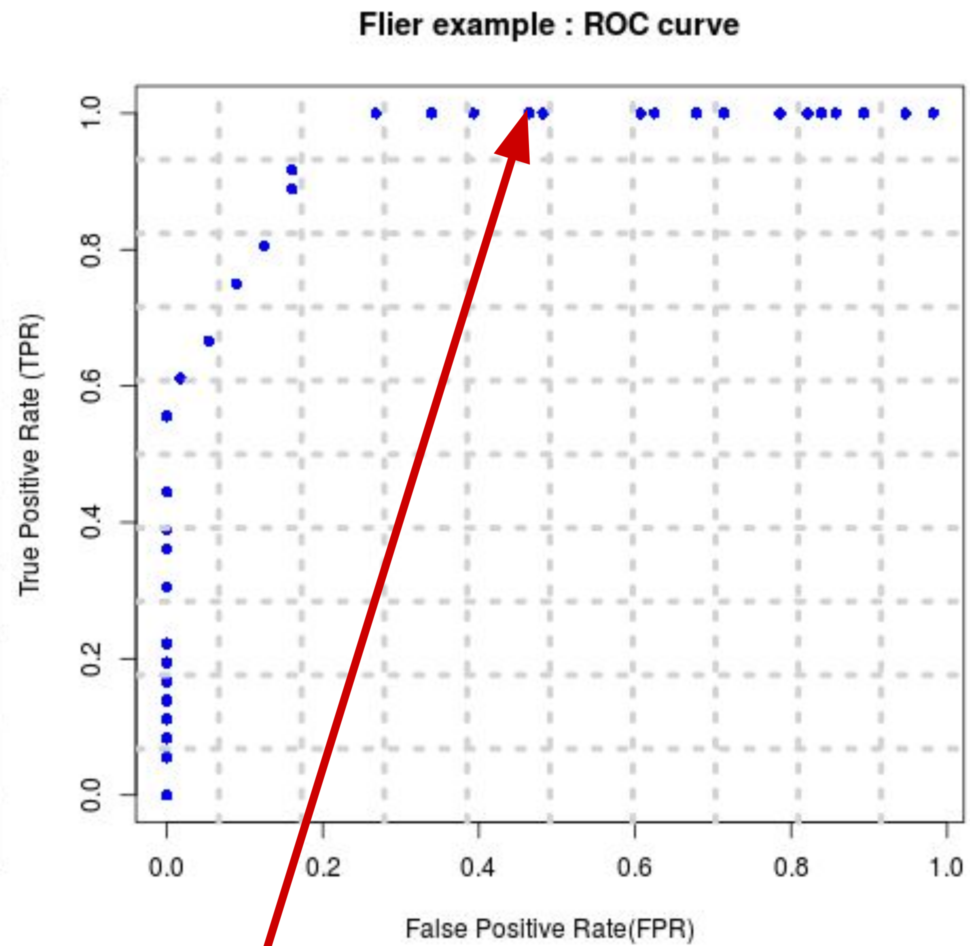
Lab activity 1 : ROC curve

1. For the flier mailing example, build a 1D classifier. The classifier has one parameter namely a threshold (on age). Plot an ROC curve as follows.

- Vary the threshold starting from a minimum value and varying in steps.
  Each threshold thus corresponds to a different parameter setting of the classifier.
- Compute the predicted class labels obtained for each threshold and hence compute True Positive Rate (Sensitivity) and False Positive Rate (1-Specificity)
- Each threshold thus corresponds to (FPR,TPR) pair, i.e. one point on the ROC curve. Compute the True Positive Rate, False Positive rate for each threshold and plot all these points on the ROC curve.
- Using the ROC plot or otherwise, compute an optimum threshold

2. Build a logistic regression model. Use the probabilities output by the model as inputs to a 1D classifier and compute an optimal threshold (on probabilities output by logistic regression) as above. Vary the threshold and plot these points on an ROC curve.





Setting the threshold at Age=40, yields a classifier that yields a TPR = 1 and FPR = 0.4642, this corresponds to one point on the ROC curve as shown.

Threshold = 40,
TPR = 1, FPR =0.4642

Model building for Example 3 – Framingham Heart Study data

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, we need to make sure the frequency of the categories are roughly the same in both training and test set.

Test/Train split

```
> # Randomly split the data into training and testing sets
> set.seed(1000)
> split = sample.split(framingham$TenYearCHD, SplitRatio = 0.70)
>
> # Split up the data using subset
> train = subset(framingham, split==TRUE)
> test = subset(framingham, split==FALSE)
> #Check the frequency of CHD in both sets
> cat(sum(train$TenYearCHD)/nrow(train),sum(test$TenYearCHD)/nrow(test))
0.1519542 0.1517296
```

CSE 7202c

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.
  See for eg:

https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

CSE 7202c

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.

Training Set

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

Testing Set

10-year CHD risk		Predicted	
Actual		True	False
	True	12	158

CSE 7202c

Gains and Lift Charts

- In some business problems, it is not good enough to just classify. For example, in direct mail or phone marketing campaigns, where it costs money to send a mail to each prospect, it is better to be able to rank the prospective buyers by their probability to buy. That way, you can order them and start calling or mailing them in their decreasing order of propensity to buy.

- **Lift** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model (random selection).

Gains and Lift Charts

- A Lift Chart describes how well a model ranks samples in a particular class.

- The greater the area between the lift curve and the baseline (random selection), the better the model.

https://www.datasciencecentral.com/profiles/blogs/understanding-and-interpreting-gain-and-lift-charts

Gains and Lift Charts

- A company sends mail catalogs to prospective buyers. It costs the company $1 to print and mail one catalog.
- From past data, they know the response rate is 5%, i.e., if 100,000 prospective customers are contacted, 5000 buy.
- This means that if there is no model and the company randomly contacts the prospects, they will have the following result.

No. of customers contacted	No. of responses
10000	500
20000	1000
30000	1500
.	.
.	.
100000	5000

CSE 7202c

Gains and Lift Charts

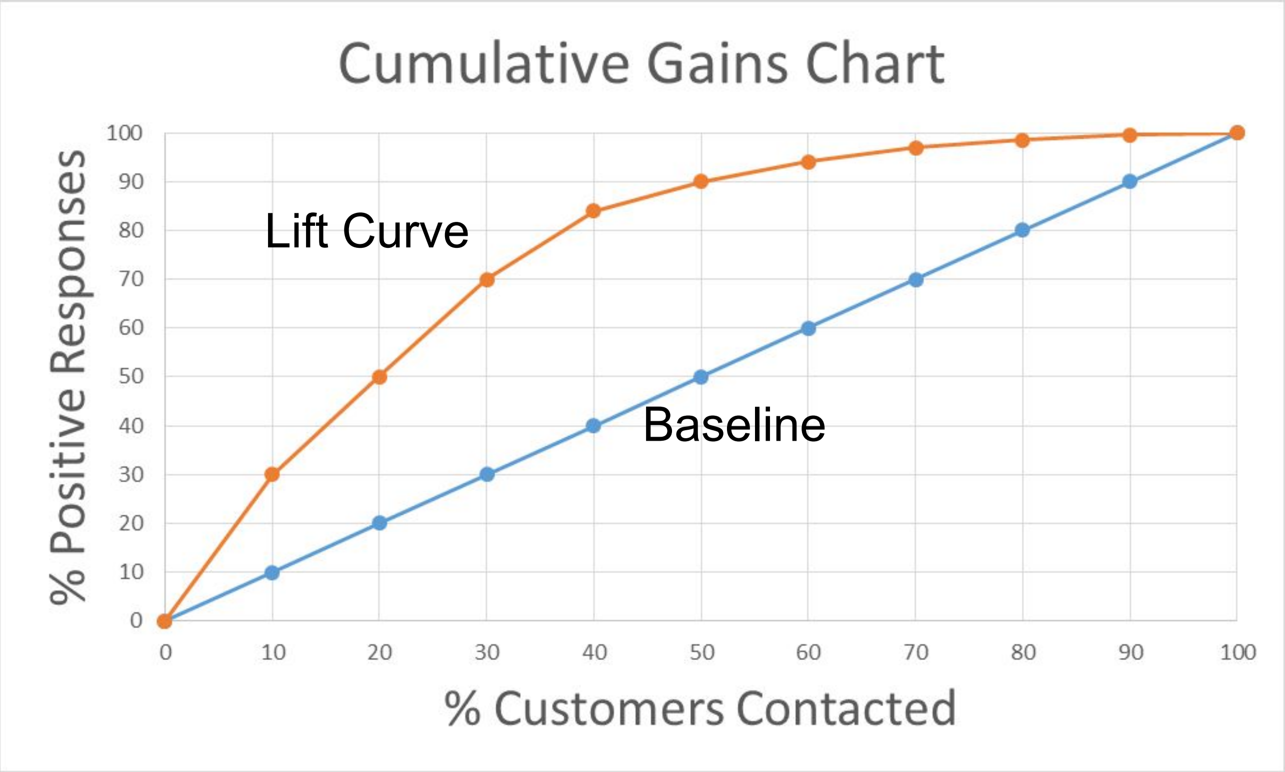
- With a predictive model, where the model assigns a probability to each customer, the customers are ordered and divided into deciles (or any other quantiles). They are then called in decreasing order of probability to buy.

Cost ($)	Decile contacted	Cumulative responses
10000	10 (top decile)	1500
20000	9	2500
30000	8	3500
40000	7	4200
50000	6	4500
60000	5	4700
70000	4	4850
80000	3	4925
90000	2	4975
100000	1	5000

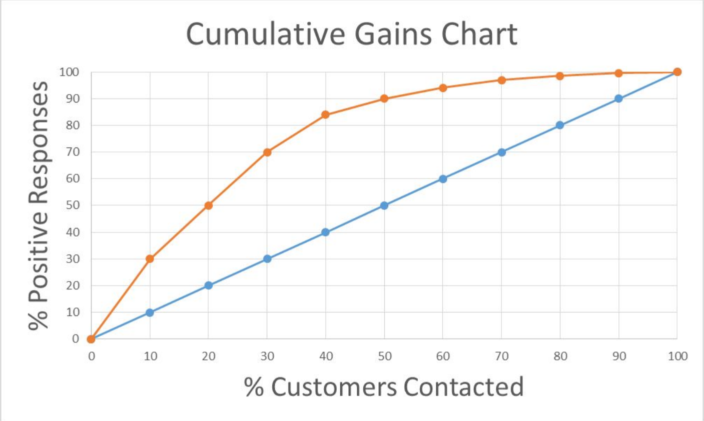
Gains and Lift Charts

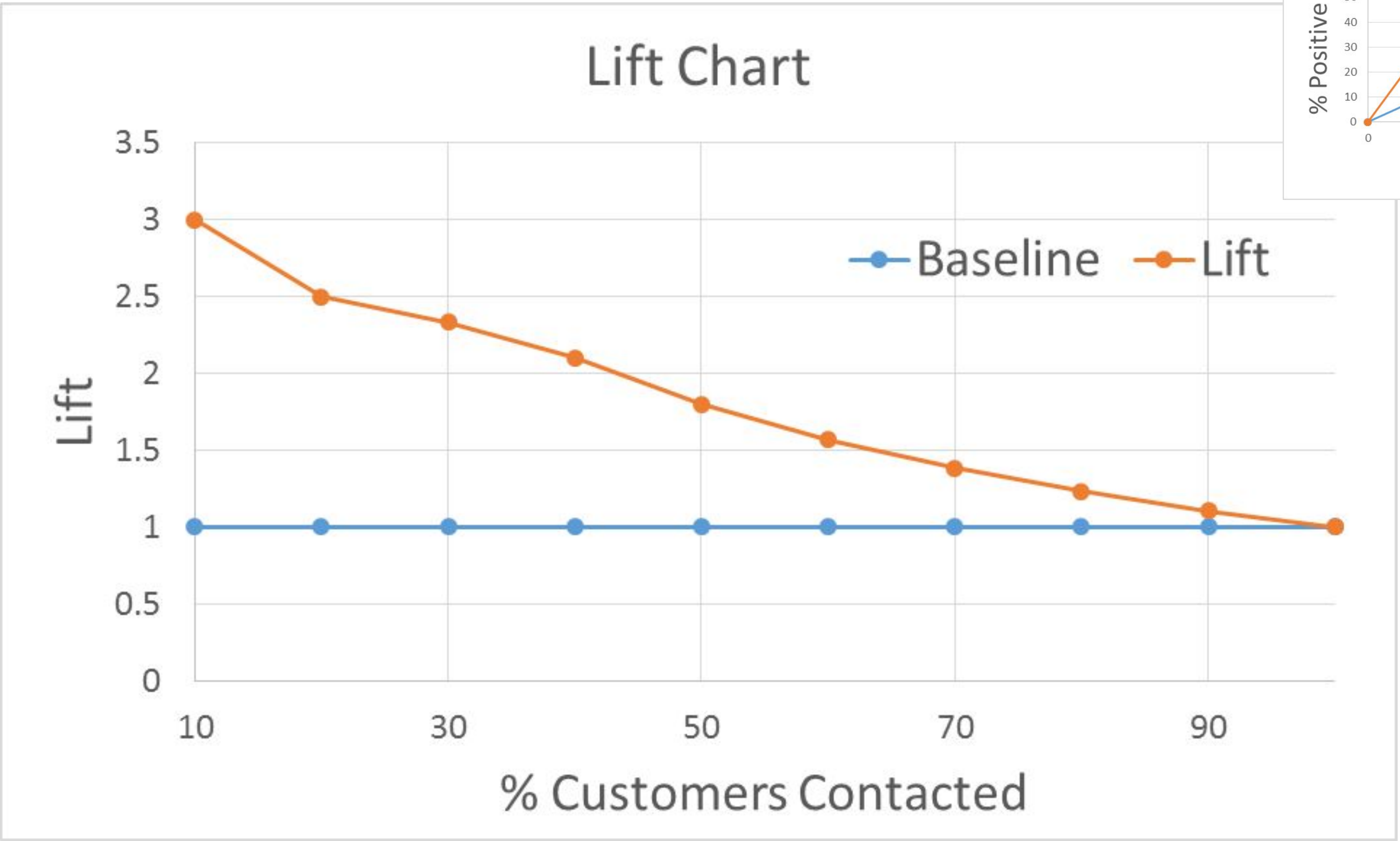
% Called	Called at Random	Called According to Model Score
0	0	0
10	10	30
20	20	50
30	30	70
40	40	84
50	50	90
60	60	94
70	70	97
80	80	98.5
90	90	99.5
100	100	100

Cost ($)	Decile contacted	Cumulative responses
10000	10 (top decile)	1500
20000	9	2500
30000	8	3500
40000	7	4200
50000	6	4500
60000	5	4700
70000	4	4850
80000	3	4925
90000	2	4975
100000	1	5000



Gains and Lift Charts





- Max lift of 3 at the top decile.
- Model advantage diminishes as more customers are contacted, especially in lower deciles.
- Useful to compare different models.

NAÏVE BAYES ALGORITHM

Classification problems

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

Naïve Bayes Algorithm

- A simple classifier that performs surprisingly well on a large class of problems
- It belongs to a class of methods called **Generative Learning Models**
- It works best when all the predictor variables are categorical variables.
- Very frequently used in text mining, character image analysis problems.

Review of Bayes Theorem

A review problem :
Suppose there are only two factories A and B that produce a particular machine component. Suppose that it is known from historical data that Factory A on average produces 3.5 defective pieces per 1000 and factory B produces 2 defective pieces per thousand.
B accounts for 60% of total production and A for the remaining.

(a) Compute the probability of a randomly chosen piece (corresponding to that machine component) being defective.
Hint : Use total probability formula.

(b) Suppose a particular piece was chosen at random and found to be defective. What is the probability that it was manufactured in factory A?
Hint : Use Bayes theorem and express aposteriori probabilities in terms of prior probabilities and likelihood.

Recall conditional probability

$$P(Y=y|X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$$

Classification according to Maximum Aposteriori Probability (MAP) rule

MAP rule : Assign the class label which corresponds to the **maximum aposteriori probability**

Maximum Aposteriori Probability (MAP) rule

Given an observation x, assign the class which yields highest value for $P(y_j|x)$ i.e.

$$k^* = argmax_j P(y_j|x)$$

If there are K classes y_1, y_2, y_K, compute $P(y_1|x), \ldots, P(y_K|x)$ and assign to x the class that yields the highest value among these.

Recall that,

$$P(y_j|x) = \frac{P(x|y_j)P(y_j)}{P(x)}$$

Classification according to Maximum Aposteriori Probability (MAP) rule and Bayes Theorem

Question. But how to compute P(class=y_j | **x**)   **j = 1,…,K?**
Answer. Use Bayes Theorem and related results.

$$P(\text{class}=y_j \mid \mathbf{x}) = \mathbf{P}(\mathbf{x} \mid \text{class} = y_j) \quad \times \quad P(\text{class} = y_j) \;\; / \;\; P(\mathbf{x})$$

Posterior probability

Class conditional probability also called **Likelihood**

Prior probability

Probability of observing x

Note that the denominator (P(**x**) is the same for all classes and is positive.
We need to focus only on numerator, if interested in just finding out which y_j yields the highest P(class=y_j/ **x)**

An example : Predict probability of playing tennis given weather conditions

Play-tennis example: estimating $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14	
P(n) = 5/14	

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 2/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

Naive Bayes example (Lab activity)

Goal: Given a weather condition, eg. **[rain, cool, normal, true]** predict whether tennis can be played?

Method : Assign the class label corresponding to the **Maximum Aposteriori Probability (MAP)** rule

i.e. Compute P(class = p | o=rain,t=cool,h=normal,w=true)
and P(class = n | o=rain,t=cool,h=normal,w=true)

Whichever probability is higher, we would assign that corresponding class label.

Naïve Bayes example (Lab activity)

The **aposteriori probabilities** are as below.

$$P(class = p|o = rain, t = cool, h = normal, w = true) =$$
$$\frac{P(o = rain, t = cool, h = normal, w = true|class = p)P(class = p)}{P(o = rain, t = cool, h = normal, w = true)}$$

$$P(class = n|o = rain, t = cool, h = normal, w = true) =$$
$$\frac{P(o = rain, t = cool, h = normal, w = true|class = n)P(class = n)}{P(o = rain, t = cool, h = normal, w = true)}$$

- Most often we are interested only in determining which aposteriori probability is higher (and not the actual value, although this could also be computed if required).
- Since the denominator is the same for both, we need to focus only on numerators.

Naïve Bayes : Lab activity

- The prior probabilities P(class = p) and P(class = n) can be easily estimated from the data.

- How to compute class conditional joint probabilities such as

$$P(o = rain, t = cool, h = normal, w = true | class = p)$$

- Rarely have sufficient data to directly estimate such joint probabilities.
- Alternative : Apply (naively) conditional independence assumption

$$P(o = rain, t = cool, h = normal, w = true | class = p) =$$

$$P(o = rain | class = p) \times$$
$$P(t = cool | class = p) \times$$
$$P(h = normal | class = p) \times$$
$$P(w = true | class = p)$$

CSE 7202c

Naïve Bayes : Lab activity

In general, if there are N attributes referred to as $x_1, x_2, \ldots, x_N$ and we wish to

compute joint probabilities such as $P(x_1 = v_1, x_2 = v_2, \ldots, x_N = v_N)$, conditional independence assumption yields

$$P(x_1 = v_1, x_2 = v2, \ldots, x_N = v_N | y_j) = P(x_1 = v_1 | y_j) P(x_2 = v_2 | y_j) \ldots P(x_K = v_K | y_j)$$

The conditional probabilities such as $P(x_1 = v_1 \mid y_j)$ can be estimated relatively easily from given data.

Naïve Bayes summary

- No parametric fit needed to compute the class
- Prior probabilities can be computed from data
- Individual conditional probabilities were evaluated, and using Bayes relationship the final class probability was evaluated

Naïve Bayes Assumption

- The key assumption of independence of features, is almost never true (and often demonstrably false)

- Still Naïve Bayes does surprisingly well in a lot of situations

Additional links

https://onlinecourses.science.psu.edu/stat504/node/149/
Logistic Regression : An online course STAT 504 offered by Penn State Eberly College of Science

https://www.coursera.org/lecture/machine-learning/classification-wlPeP
Andrew Ng's Coursera Machine Learning course (topic Classification)

https://www.youtube.com/watch?v=zAULhNrnuL4 Statistics 101: Logistic Regression, An Introduction, Video lecture by Brandon Foltz

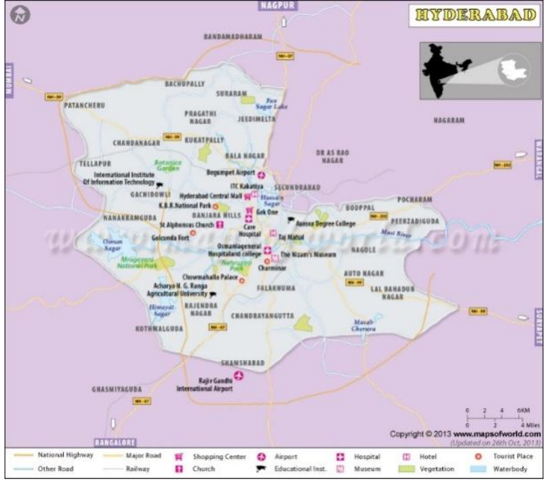
https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Logistic Regression for Machine Learning, Jason Brownlee

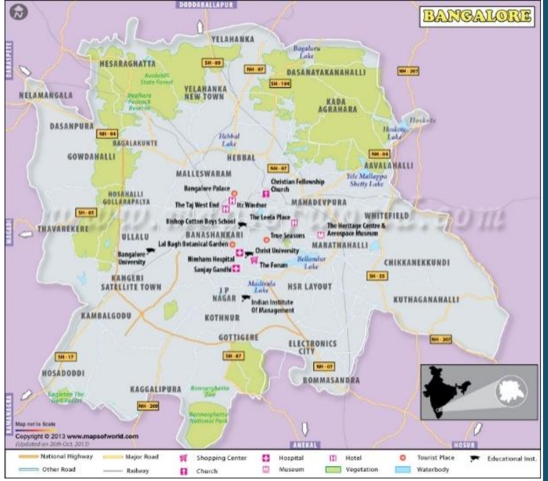
https://en.wikipedia.org/wiki/Receiver_operating_characteristic Receiver operating characteristic

https://www.datasciencecentral.com/profiles/blogs/understanding-and-interpreting-gain-and-lift-charts Understanding And Interpreting Gain And Lift Charts

https://machinelearningmastery.com/naive-bayes-for-machine-learning/Naive Bayes for Machine Learning, Jason Brownlee



Inspire...Educate...Transform.



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

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

Social Media

Web:	
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	
SlideShare:	
LinkedIn:	

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.