Logistic Regression

Heather Berginc (Creditt: Frank Burkholder, A.Richards)

Success Criteria

Today I will be successful if I can…

- Decide which class to make positive in binary classification
- Decide how classifications are made using predicted probabilities and a threshold
- Describe how logistic regression determines probabilities
- Interpret logistic regression coefficients
- Calculate TP, FP, FN and TN
- Construct a confusion matrix
- Describe what an ROC curve is, and how it describes the performance of a classifier

Review

Identify the following scenarios as classification or regression. Prepare to explain the target for each as well.

1. Investigating how much TV advertising impacts sales for each quarter.
2. Medical researchers attempt to understand the relationship between drug dosage and blood pressure.
3. Detecting spam which should be delivered to your junk folder.
4. Investigating and deciding whether or not someone is qualified for a loan.
5. Measuring the effect of fertilizer and water on crop yields.
6. Measuring the effect of different training regimens have on a player's performance.
7. You'd like to predict the winner of the Kentucky Derby.

Regression vs Classification

Considering determining the winner of the Kentucky Derby...

How would you approach this problem?

Regression vs. Classification

You could calculate either of these numerical values for each horse and sort:

- Finish Time
- Lengths behind winner

Regression vs. Classification

You could calculate either of these numerical values for each horse and sort:

- Finish Time
- Lengths behind winner

Binary Classification

- Win or Lose

Multinomial Classification

- First, Second, Third, Forth, etc.

Regression vs. Classification

You could calculate either of these numerical values for each horse and sort:

- Finish Time
- Lengths behind winner

Binary Classification

- Win or Lose

Multinomial Classification

- First, Second, Third, Forth, etc.

But what about:

- 1,2,3,4 etc. Regression? (does -1, or 3.5 make sense?)
- 1,2,3,4 etc. Classification? (1 better than 2, 2 better than 3, etc.)

In Binary Classification, picking the + class

- In binary classification, there are two classes, e.g.:
  - Negative, positive
  - False, true
  - No default, default
  - No cancer, cancer
  - No churn, churn
- These options are encoded into an indicator variable taking value of 0 or 1.
- Usually you care about one of them more than the other, and that is the one you make 1, the **positive class**
  - Default
  - Cancer
  - Churn
- This matters when you pick an **evaluation metric** for your model (later)

Deciding whether a data point is the + class

A sklearn classification model (like logistic regression) will predict which class each datapoint is in:

```
>>> model = LogisticRegression()
>>> model.fit(X_train, y_train)
>>> y_hat = model.predict(X_test)
>>> y_hat
    array([[0],
           [1],
           [0],
            …,
           [1]])
```

But - and this is important - **this is not actually that useful.**

Deciding whether a data point is the + class

The *probability* that each row of data belongs to the positive class - that's much more useful:

```
>>> model = LogisticRegression()
>>> model.fit(X_train, y_train)
>>> y_hat_probs = model.predict_proba(X_test)[:,1]
>>> y_hat_probs
    array([[0.33],
           [0.78],
           [0.05],
              …,
           [0.94]])
```

Now, you can pick a probability threshold, say 0.4, that you can use to transform the probabilities into classifications:

```
>>> threshold = 0.4
>>> y_hat = (y_hat_probs >= threshold).astype(int)
```

Breakout 1

Use the predicted probability of the + class and the given threshold to fill in the predictions in each of the tables below. One is filled out for you.

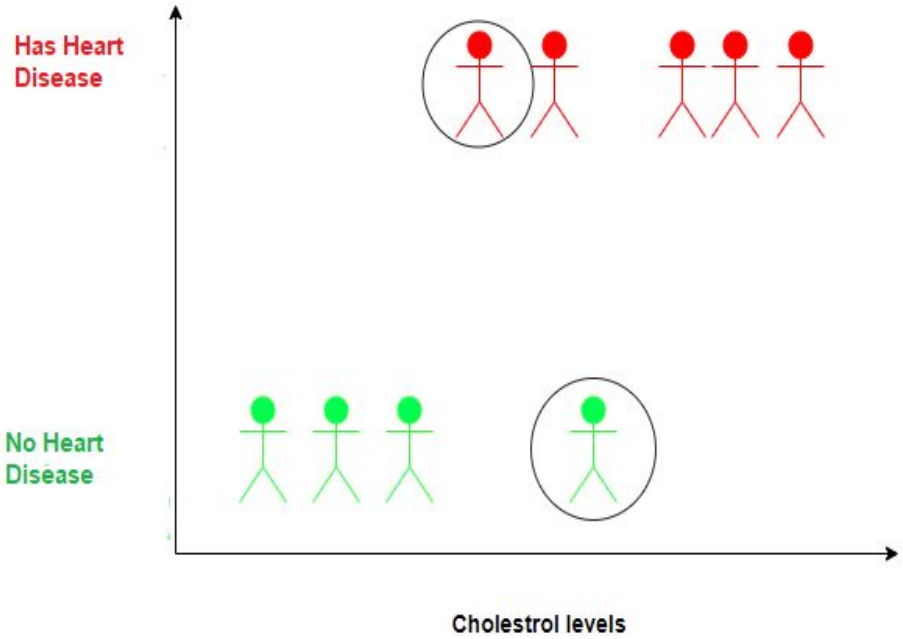
threshold:	1.0
y_hat_prob	y_hat
0.67	
0.20	
0.98	
0.03	

threshold:	0.75
y_hat_prob	y_hat
0.67	
0.20	
0.98	
0.03	

threshold:	0.5
y_hat_prob	y_hat
0.67	1
0.20	0
0.98	1
0.03	0

threshold:	0.25
y_hat_prob	y_hat
0.67	
0.20	
0.98	
0.03	

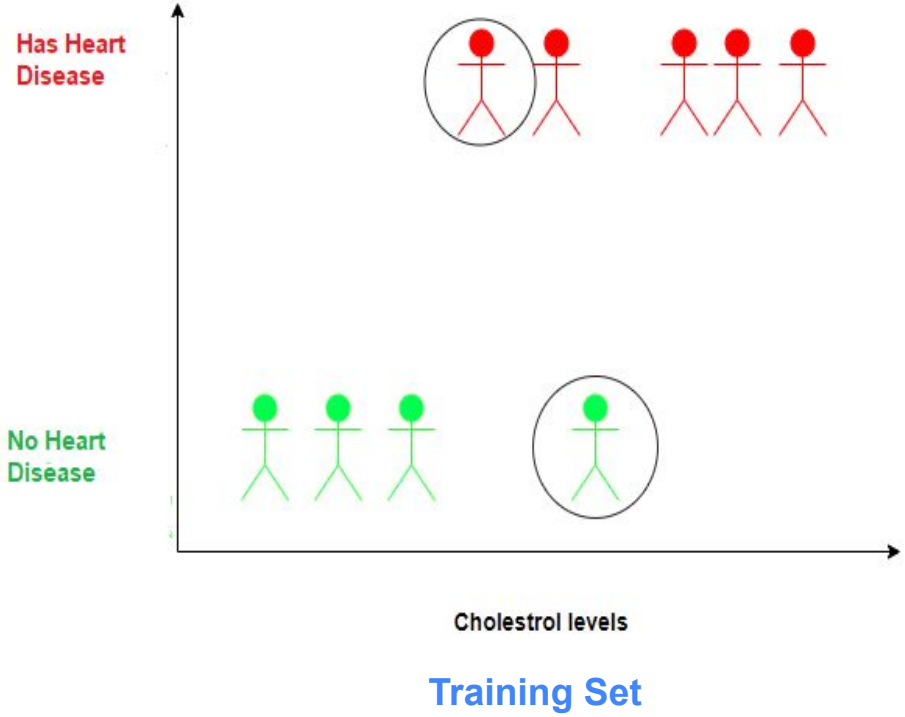
The idea behind the Logistic Function



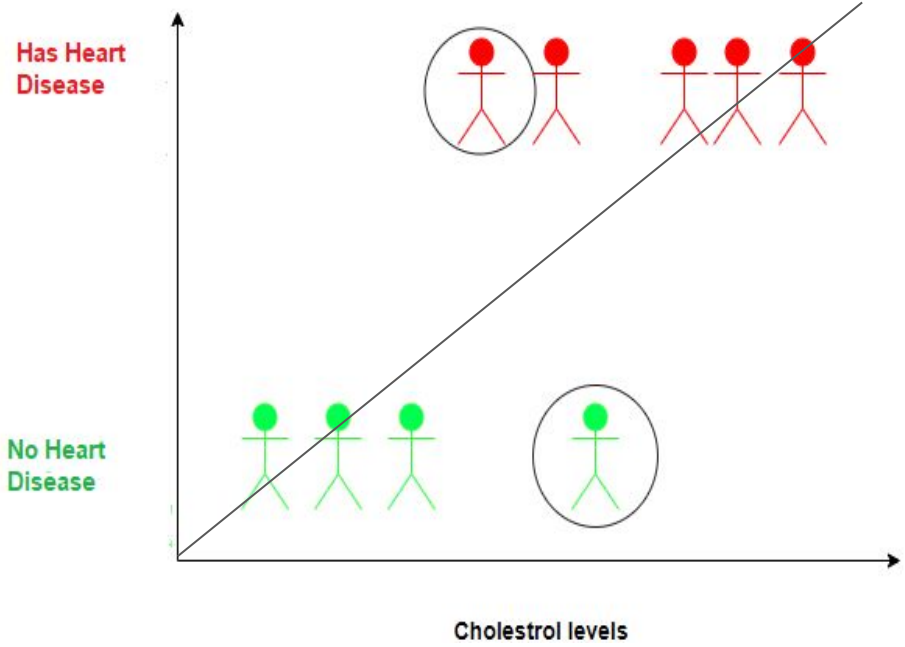
Training Set

Images: Parul Pandey, TowardsDataScience.com

The idea behind the Logistic Function



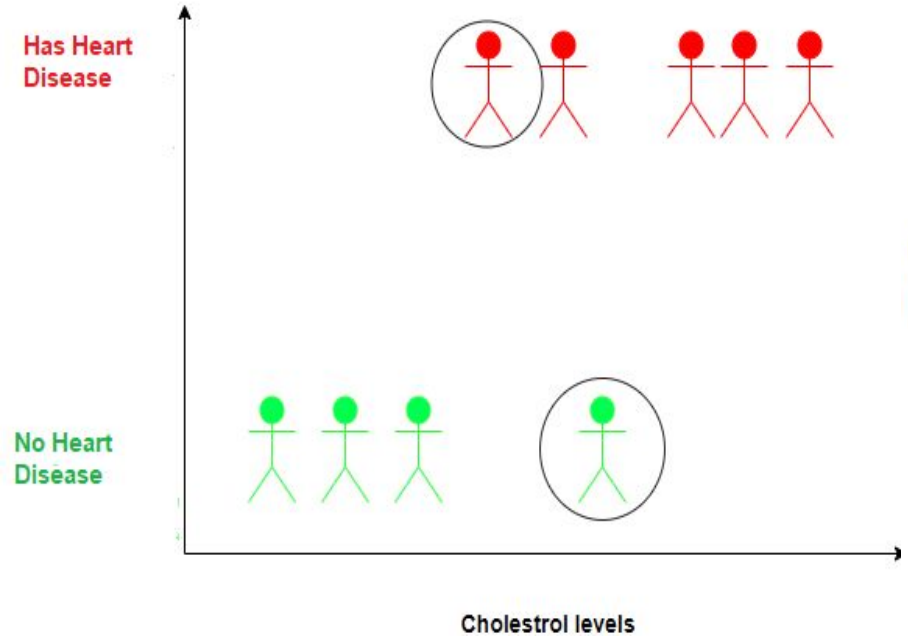
Training Set



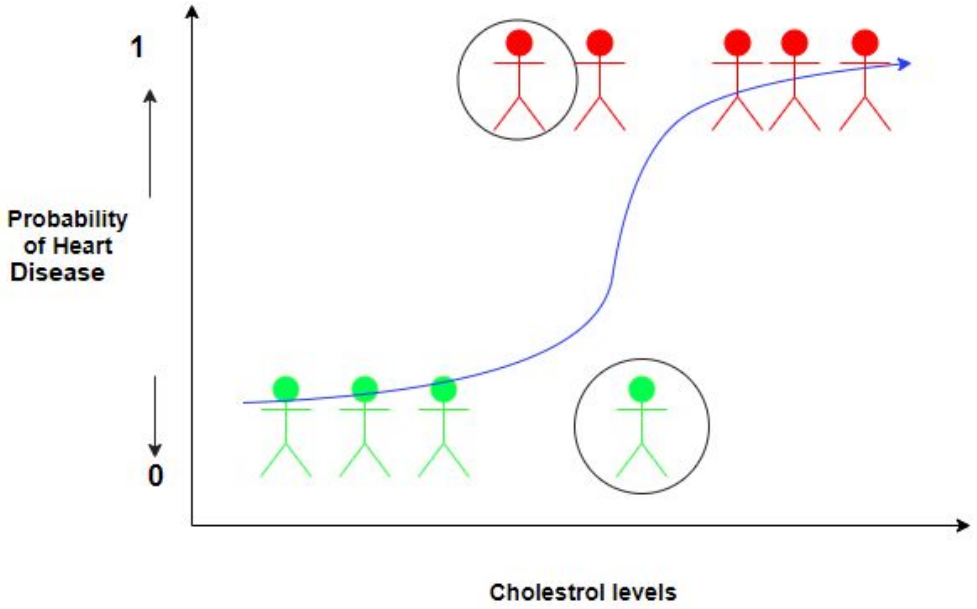
Linear Fit (BAD)

Images: Parul Pandey, TowardsDataScience.com

The idea behind the Logistic Function



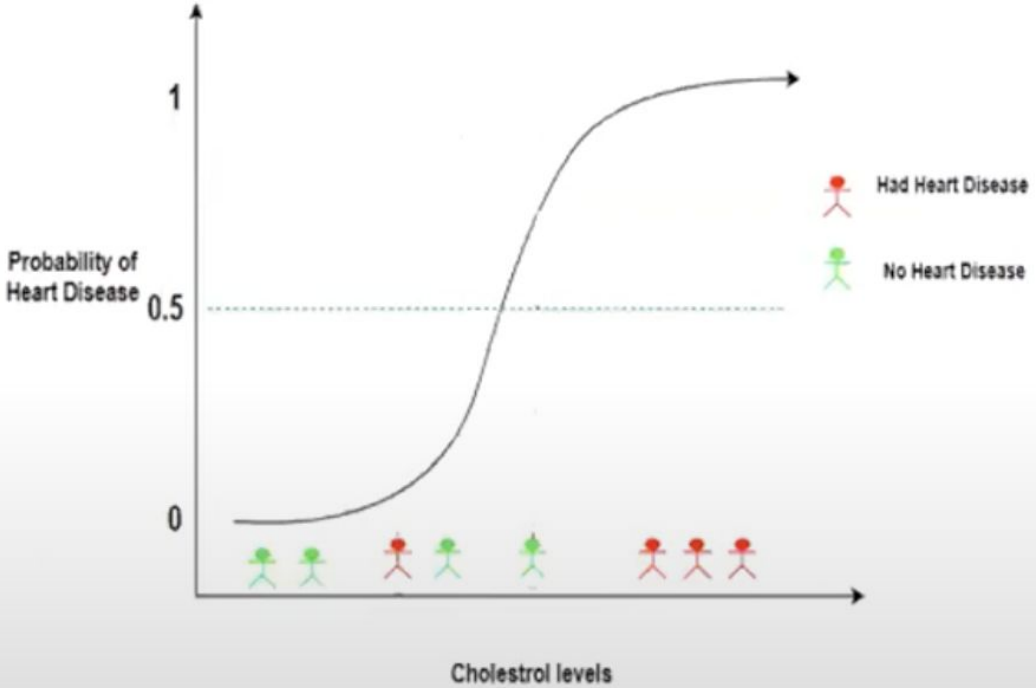
Training Set



Logistic/Sigmoid Fit

Images: Parul Pandey, TowardsDataScience.com

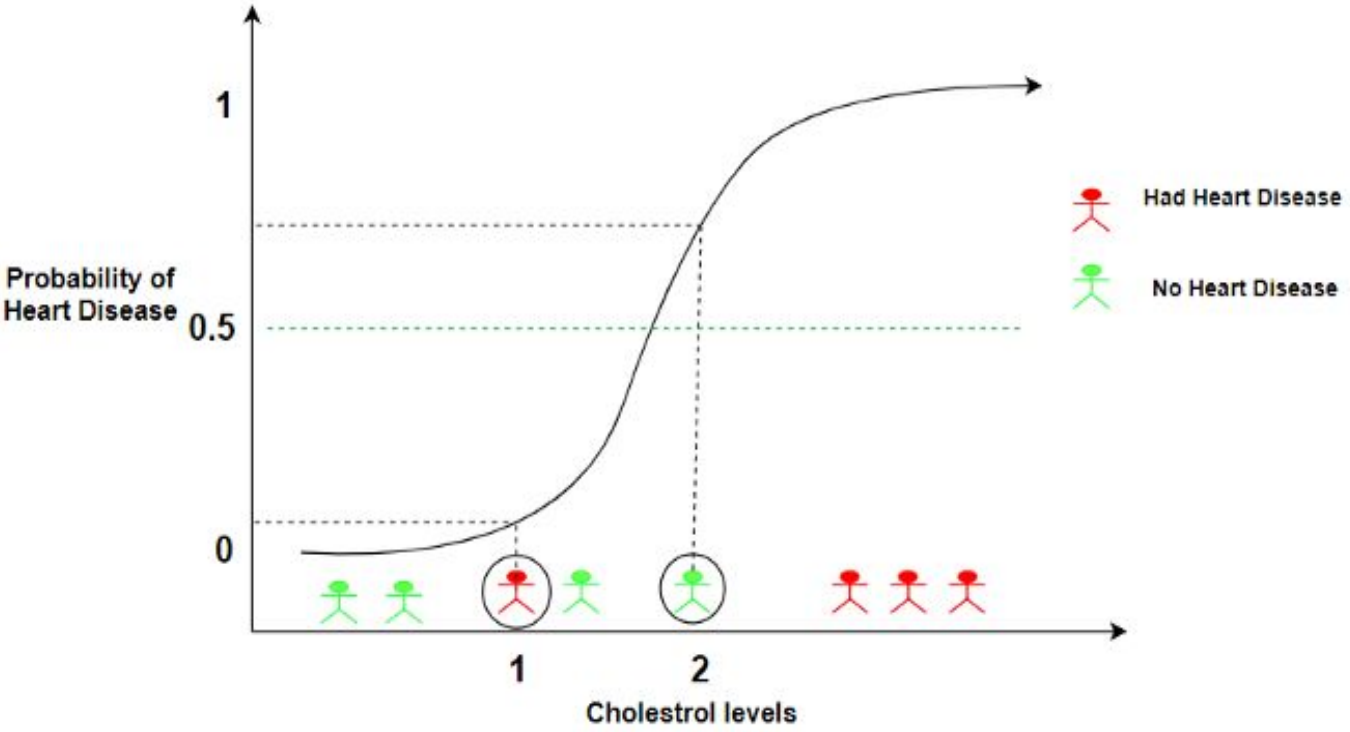
The Logistic (Sigmoid) Function



Images: Parul Pandey, TowardsDataScience.com

- Using a threshold of .5, how many people are misclassified in this model?
- If we wanted to make sure we caught everyone who has heart disease, what should we change the threshold to?

The Logistic (Sigmoid) Function



- Using a threshold of .5, how many people are misclassified in this model?
- If we wanted to make sure we caught everyone who has heart disease, what should we change the threshold to?

Images: Parul Pandey, TowardsDataScience.com

Determining probabilities in logistic regression

- Question: if logistic regression does classification, why is it called regression?

both can take value from -∞ to + ∞

$$? = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Linear regression**

$$? = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Logistic regression**

Determining probabilities in logistic regression

- Question: if logistic regression does classification, why is it called regression?

- Answer: It's a model, linear in its coefficients, that regresses values on to a numerical quantity that scales from -infinity to +infinity (and that's exactly what linear regression does).

both can take value from -∞ to + ∞

$? = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$ **Linear regression**

$? = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$ **Logistic regression**

Determining probabilities in logistic regression

- In Linear Regression, the sum of the products of the coefficients and features is the target, $\hat{y}$
- In Logistic regression, the sum of the products of the coefficients and features is the target, the *log odds*

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Linear regression**

Logit function!

$$ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Logistic regression**

So we have two linear functions, but the value of the one below is hard to interpret, so we should transform this into probabilities (like previous slides).

Before we go any further: Odds

Statistical odds are the long run ratio of the probability of an event occurring to it not occurring.

For example, the odds of rolling a 3 on a fair 6 sided die are 1 to 5.

$$odds\ of\ \ 3 = 1\ to\ 5 = \frac{1}{5} = \frac{p_3}{1 - p_3}$$

$$p_3 = \frac{1}{6}$$

Determining probabilities in logistic regression

The log odds, otherwise known as the logit, links values that range from (- ∞,∞) to a probability in the range [0, 1].

[-∞, ∞] $ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$ *logistic regression*

[0, ∞] $odds = \frac{p}{1-p}$ *prob. of + class / prob. of - class*

[0, 1] p *prob. of + class*

Rearranging the logistic regression equation (1 of 4)

Solve for p

$$ln(odds) = \beta_0 + \beta_1 x_1 + \ldots$$

logistic regression

$$odds = e^{\beta_0 + \beta_1 x_1 + \ldots}$$

exponentiate both sides

Rearranging the logistic regression equation (1 of 4)

Solve for p

$$ln(odds) = \beta_0 + \beta_1 x_1 + \ldots$$

$$odds = e^{\beta_0 + \beta_1 x_1 + \ldots}$$

$$\log_b a = c \quad \text{is} \quad b^c = a$$

Rearranging the logistic regression equation (2 of 4)

Solve for p

$$odds = e^{\beta_0+\beta_1 x_1+\ldots}$$

odds

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\ldots}$$

substitute p in odds

$$p = e^{\beta_0+\beta_1 x_1+\ldots} - e^{\beta_0+\beta_1 x_1+\ldots} p$$

multiply both sides by 1 - p

Rearranging the logistic regression equation (3 of 4)

Solve for p

$$p = e^{\beta_0+\beta_1 x_1+\ldots} - e^{\beta_0+\beta_1 x_1+\ldots} p$$

multiply both sides by 1 - p

$$p + e^{\beta_0+\beta_1 x_1+\ldots} p = e^{\beta_0+\beta_1 x_1+\ldots}$$

get p on the same side

$$p(1 + e^{\beta_0+\beta_1 x_1+\ldots}) = e^{\beta_0+\beta_1 x_1+\ldots}$$

factor out p

Rearranging the logistic regression equation (4 of 4)

Solve for p

$$p(1 + e^{\beta_0+\beta_1 x_1+\ldots}) = e^{\beta_0+\beta_1 x_1+\ldots}$$

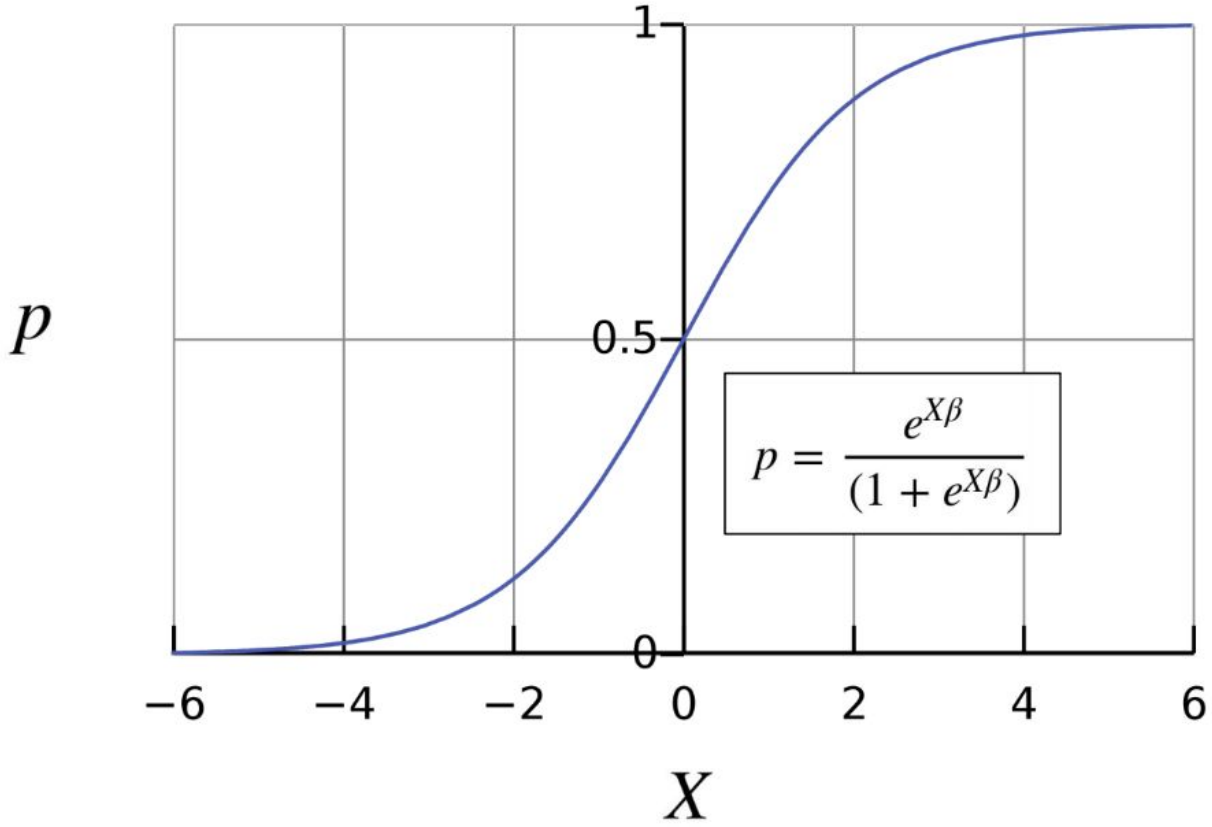
factor out p

$$p = \frac{e^{\beta_0+\beta_1 x_1+\ldots}}{(1 + e^{\beta_0+\beta_1 x_1+\ldots})}$$

divide by term to isolate p

Whew! Does this look familiar?
It's the logistic (aka sigmoid) function.

Logistic function - determining probability given $X\beta$

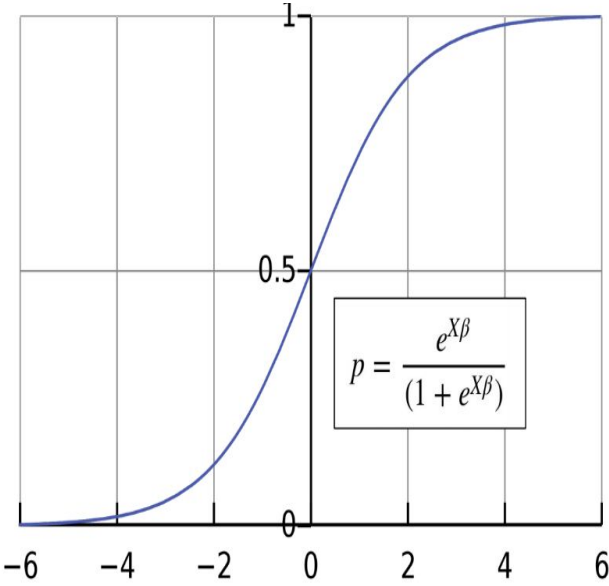


Note: XB is generally a concise way of writing $B_0 + B_1 X_1 + \ldots$

source: Wikipedia

Determining probabilities in logistic regression - RECAP

- In logistic regression prediction, XB is computed and then placed in the logistic function to determine the probability that each row of data belongs to the positive class.
- The logistic function naturally bounds the probability between 0 and 1.
- In logistic regression training, maximum likelihood is use to find the coefficients that maximize the likelihood of the existing classifications given the data X B



You'll solve for these coefficients in the Gradient Descent assignment coming up later

Interpreting logistic regression coefficients

How to interpret coefficient β_1 ?

-

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Linear regression**

-

$$ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Logistic regression**

Interpreting logistic regression coefficients

How to interpret coefficient β_1 ?

- In linear regression, for a 1 unit increase in x_1 the response $\hat{y}$ will increase by β_1 assuming all other values are held constant.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Linear regression**

- In logistic regression, for a 1 unit increase in x_1 the *log odds* will increase by β_1 assuming all other values are held constant.

$$ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$ **Logistic regression**

"... the log odds will increase …" ← True, but what does that mean? Speak of odds instead...

Interpreting logistic regression coefficients

How to interpret coefficient β_1 ?

Recall:

$$odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$$

Using the exponential power rule:

$$odds = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2}$$

If everything else held constant:

$$odds = C \cdot e^{\beta_1 x_1}$$

In logistic regression, for a 1 unit increase in X_1 the odds will increase by a factor of e^{B_1} assuming all other values are held constant.

We can calculate the growth factors!

- If B_1 is ≥1, what impact does that have on your predicted probability?

$\beta = 3$ $\quad e^3 = 20.09$

- If B_1 is negative, what impact does that have on your predicted probability?

$\beta = -2$ $\quad e^{-2} = \frac{1}{e^2} = .135$

- If B_1 is in the range [0, 1)

$\beta = .5$ $\quad e^{.5} = 1.65$

In logistic regression, for a 1 unit increase in x_1 the *odds* will increase by a factor of e^{β_1} assuming all other values are held constant.

Evaluating logistic regression (and other classification models)

- In a test or hold-out set, you should have data that has true classifications (y_{true}) that you can compare to your predictions ($\hat{y}$)
- Comparing predictions to true values, you can label each prediction a True Positive (TP), a False Positive (FP), a False Negative (FN), and a True Negative (TN):

y_{true}	$\hat{y}$	label
1	1	**TP**
0	1	**FP**
1	0	**FN**
0	0	**TN**

Evaluating logistic regression (and other classification models)

- The counts of the number of TP, FP, FN, and TN are typically summarized in a table called, appropriately, a confusion matrix.

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

- Useful metrics can be calculated from the counts in the confusion matrix.

Classification notation and metrics

condition positive (P)

the number of real positive cases in the data

condition negative (N)

the number of real negative cases in the data

---

true positive (TP)

eqv. with hit

true negative (TN)

eqv. with correct rejection

false positive (FP)

eqv. with false alarm, Type I error

false negative (FN)

eqv. with miss, Type II error

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

Metrics, a closer look

- **Accuracy**:

Of all of the data we labeled, how many did we get correct?

(TP+TN) / (TP+TN+FP+FN)

- **Precision** (Positive Predictive Value):

TP / (TP+FP)

Of all the data we labeled with a positive value, how many should have been positive?

- **Recall** (True Positive Rate):

TP / (TP+FN)

Of all of the ACTUAL positive data, how many did we correctly capture with a prediction of positive?

- **F1 Score**:

2*TP / (2*TP+FP+FN)

  - Weighted average of Precision and recall
  - Not as “verbally” interpretable

Breakout 2

Situation	What are trying to minimize? FP or FN? Both?	Therefore you should maximize? (Accuracy, Precision, Recall, F1 score)
Checking for a disease		
Detecting spam in your email		
Seal of quality test result for parachute manufacturing		
Identifying people to target with a marketing campaign		
Deciding to convict someone of a crime		

Breakout 3

Use the true value, the predicted probability of the + class and the given threshold to fill in the predictions in the tables below. Then, construct a confusion matrix and calculate the accuracy, precision, and recall.
An example is filled out for you.

thresh:	**0.5**		
y_h_prob	y_hat	y_true	label
0.67	1	0	FP
0.20	0	1	FN
0.98	1	1	TP
0.03	0	0	TN

		True	
		1	**0**
Predicted	**1**	1 (TP)	1 (FP)
	0	1 (FN)	1 (TN)

Accuracy = 2/4 = 0.5
Precision = ½ = 0.5
Recall = ½ = 0.5

Example

thresh:	**0.20**		
y_h_prob	y_hat	y_true	label
0.67		0	
0.20		1	
0.98		1	
0.03		0	

		True	
		1	**0**
Predicted	**1**		
	0		

Accuracy =
Precision =
Recall =

Determining the overall performance of a classifier

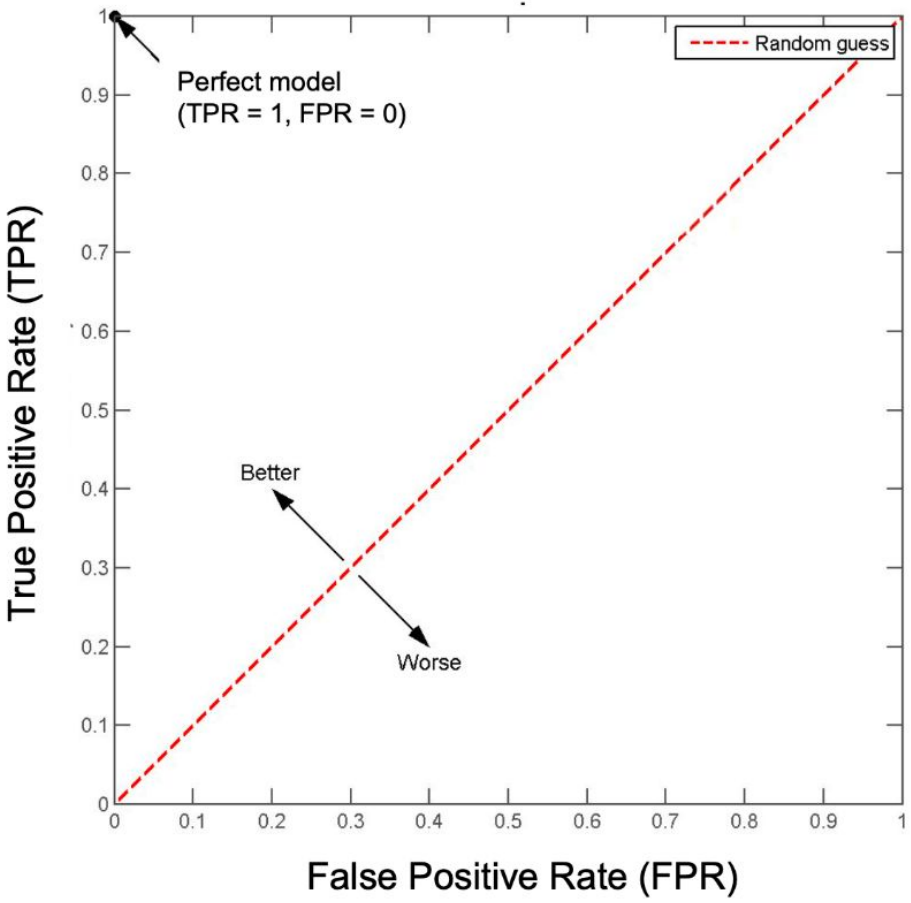
- You may have noticed that you get different model performance metrics, depending on the threshold you use to determine the classifications from the prediction probabilities.

- This is obviously not idea - would like to have a metric that incorporates all possible values of the threshold.

- There is a curve and metric that encompass this performance: the Receiver Operating Characteristic (ROC curve).

Receiver Operating Characteristic (ROC) curve

- ROC curve plots the TPR vs the FPR for all thresholds of interest.
- It's easy to get a good TPR with a high FPR (just guessing the positive class all the time).
  - very low threshold usually gives many FPs
- It's difficult to get a good TPR with a low FPR.
  - very high threshold minimizes FPs, but usually miss some TPs
- Total performance quantified by Area Under the Curve (AUC)
  - Perfect: AUC = 1
  - Random guessing: AUC = 0.5

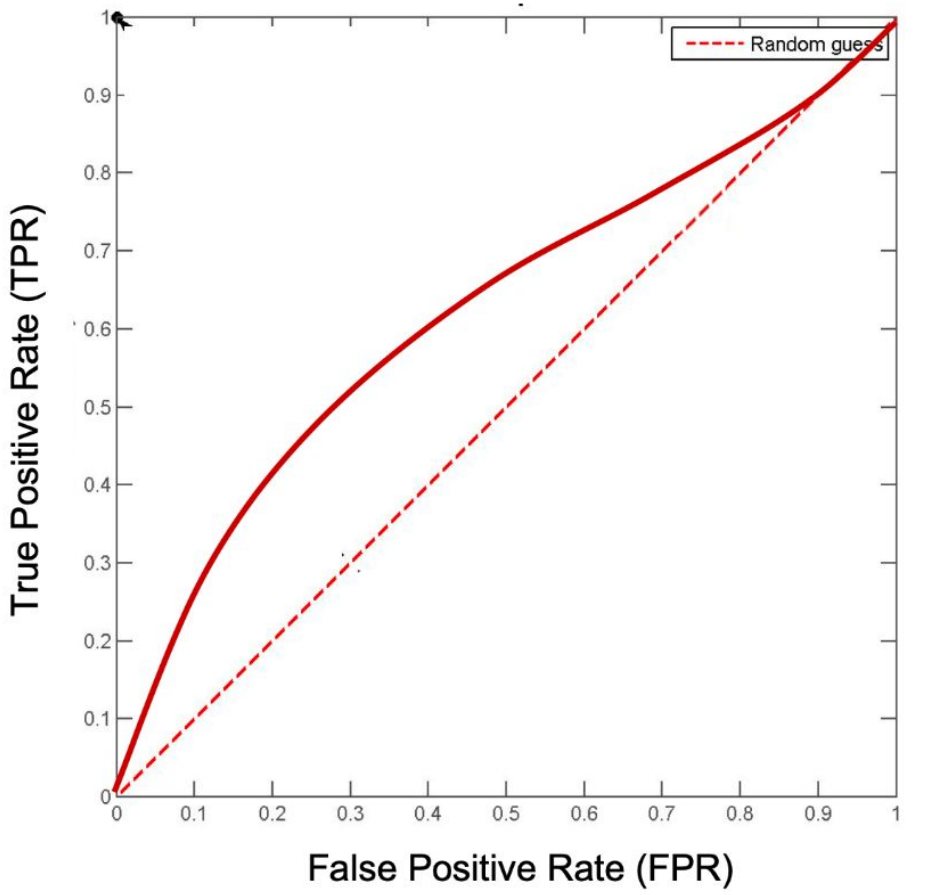


Source: Wikipedia

Receiver Operating Characteristic (ROC) curve

A so-so model

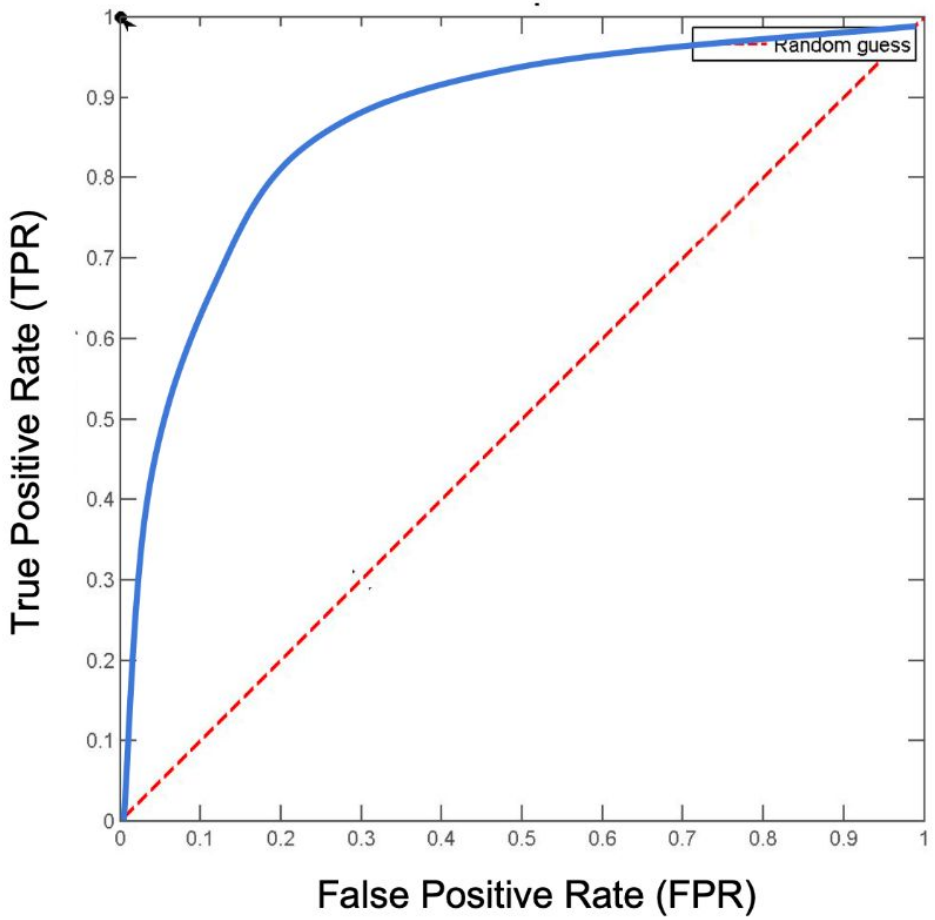
AUC ~ 0.65



Receiver Operating Characteristic (ROC) curve

A good model

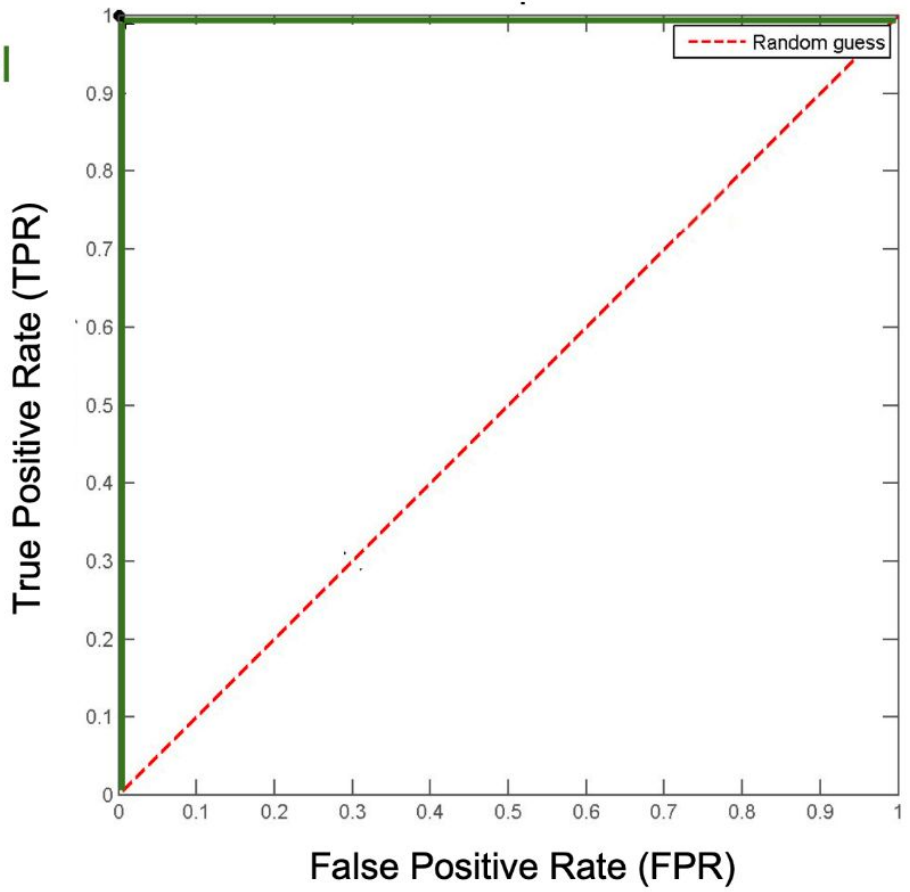
AUC ~ 0.8



Receiver Operating Characteristic (ROC) curve

An unreasonably good model

AUC ~ 0.999999



Constructing a ROC curve

1. Fit a classifier to your training data and get positive class prediction probabilities.
2. Sort these probabilities - can be low to high or high to low. These probabilities are doubling as your thresholds.
3. Compare all your predicted probabilities to each threshold to get predictions and from prediction and true values, fill out a confusion matrix. You get a different matrix for each threshold.
4. From the confusion matrix, calculate TPR and FPR
5. The FPR and TPR for a given threshold and confusion matrix are an (x,y) pair for the ROC - plot them!

Quick Demo

horse_or_dog.ipynb

Review Success Criteria

Today I will be successful if I can…

- Decide which class to make positive in binary classification
- Decide how classifications are made using predicted probabilities and a threshold
- Describe how logistic regression determines probabilities
- Interpret logistic regression coefficients
- Calculate TP, FP, FN and TN
- Construct a confusion matrix
- Describe what an ROC curve is, and how it describes the performance of a classifier