GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 9: LOGISTIC REGRESSION

LAST LESSONS

LINEAR REGREGSSION

POLYNOMIAL REGRESSION

REGULARIZATION

FEATURE CREATION

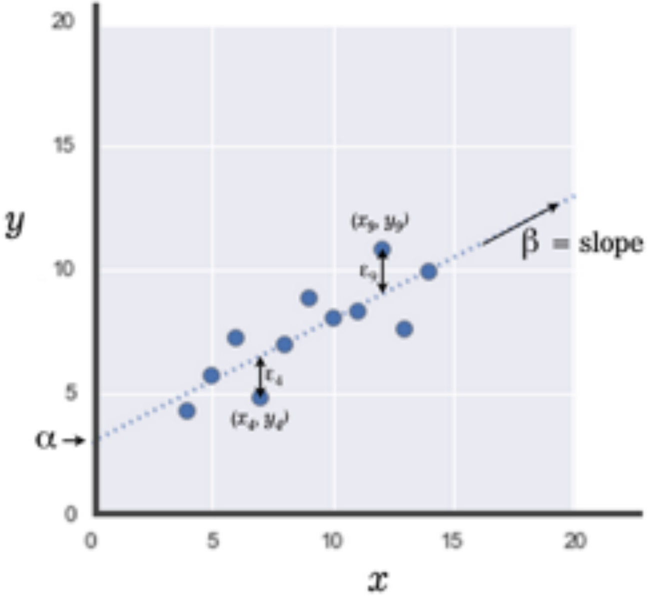
ASSIGNMENT #2: SALARY PREDICTION

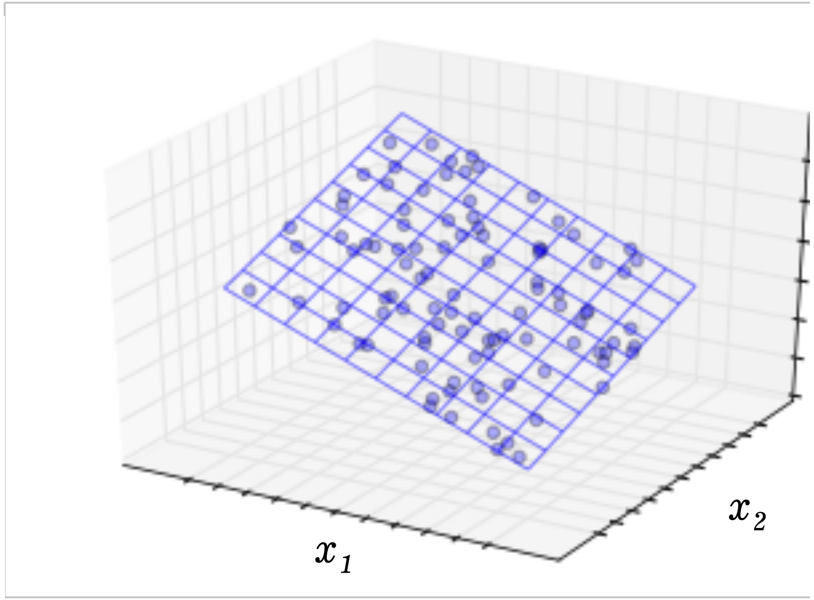
any questions?

I. REGRESSION RECAP
II. LOGISTIC REGRESSION
III. INTERPRETING RESULTS
IV. DECISION BOUNDARIES
V. EVALUATING CLASSIFIERS

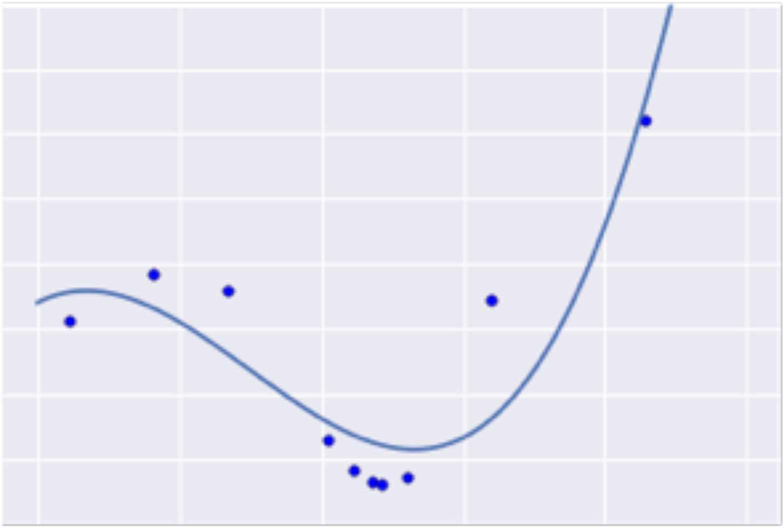
I. REGRESSION RECAP

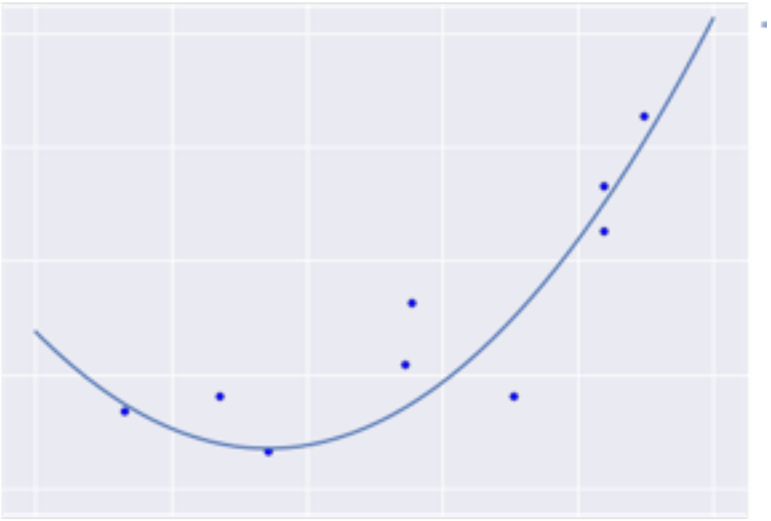
$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$





$$y = \alpha + \beta_1 x_1 + \beta_1 x_1{}^2 + \ldots + \varepsilon$$



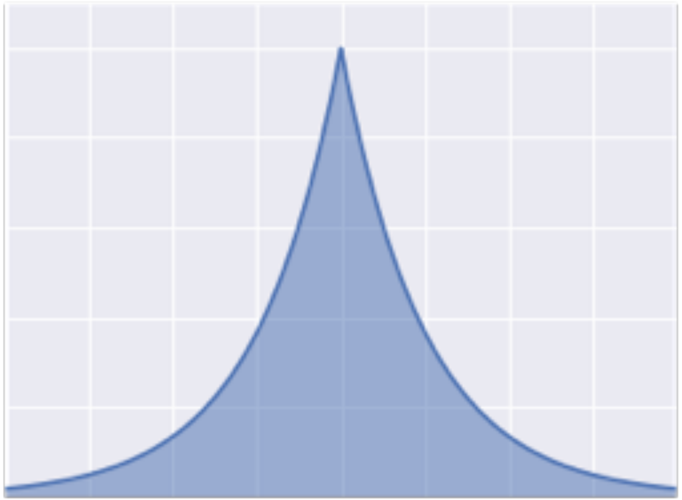


OLS: $\min (\|y - x\beta\|^2)$

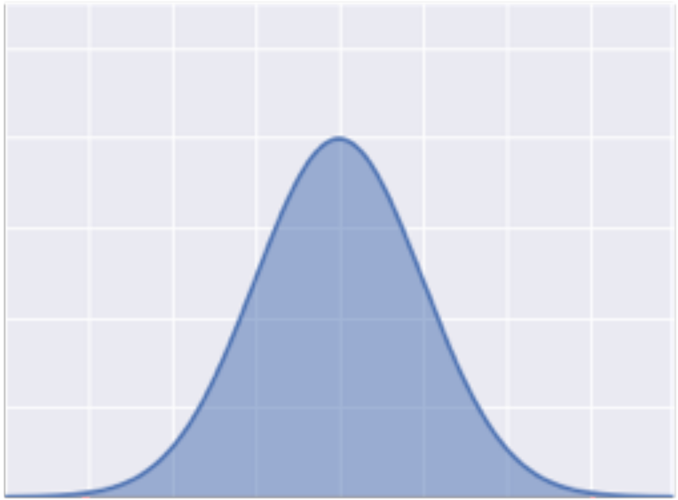
Lasso (L1): $\min (\|y - x\beta\|^2 + \lambda\|\beta\|)$

Ridge (L2): $\min (\|y - x\beta\|^2 + \lambda\|\beta\|^2)$

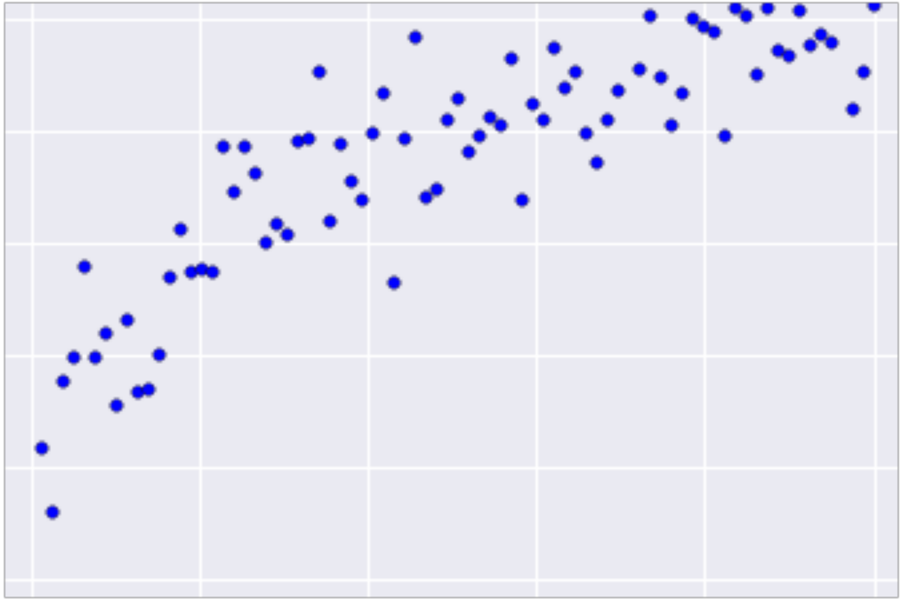
Laplace distribution



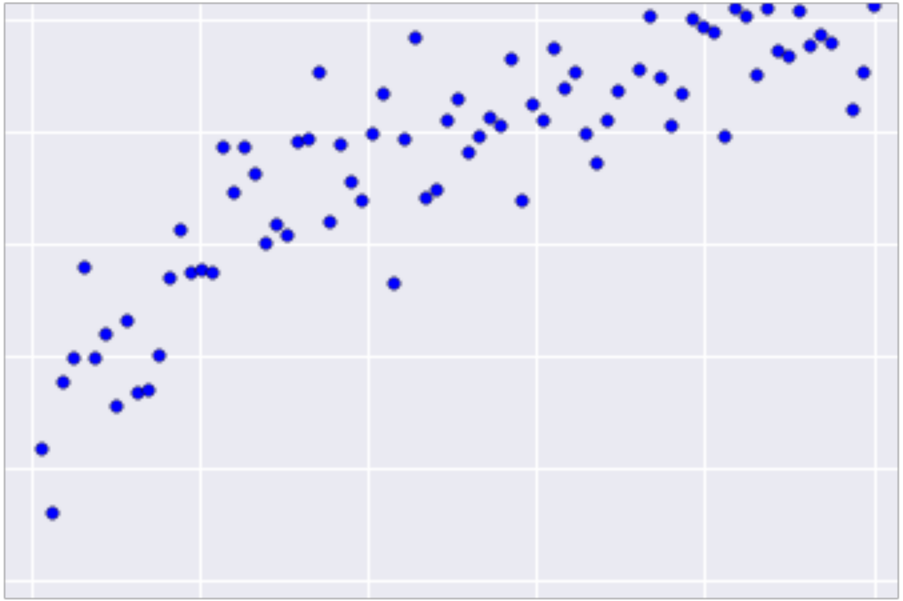
Gaussian distribution



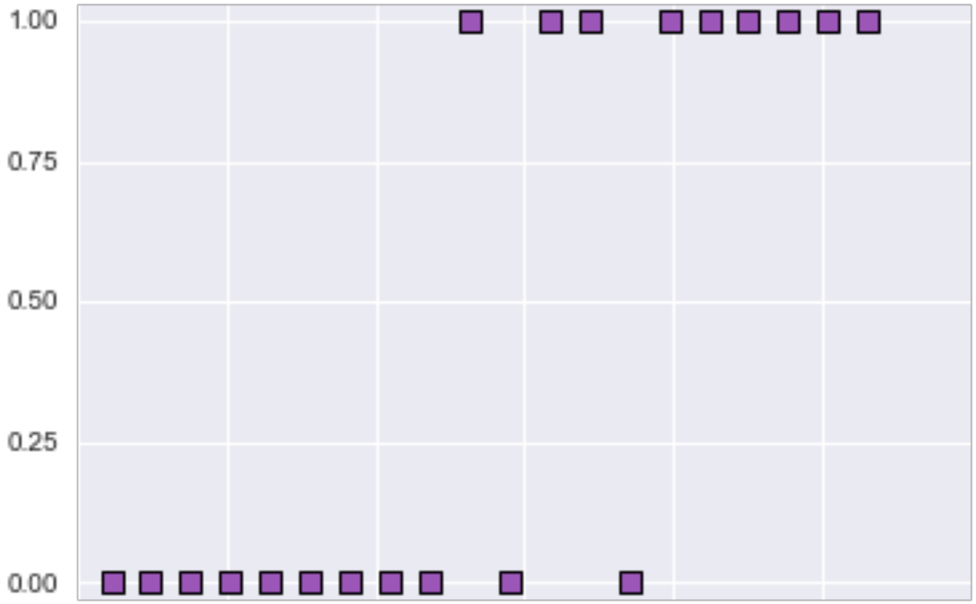
$$y^2 = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$



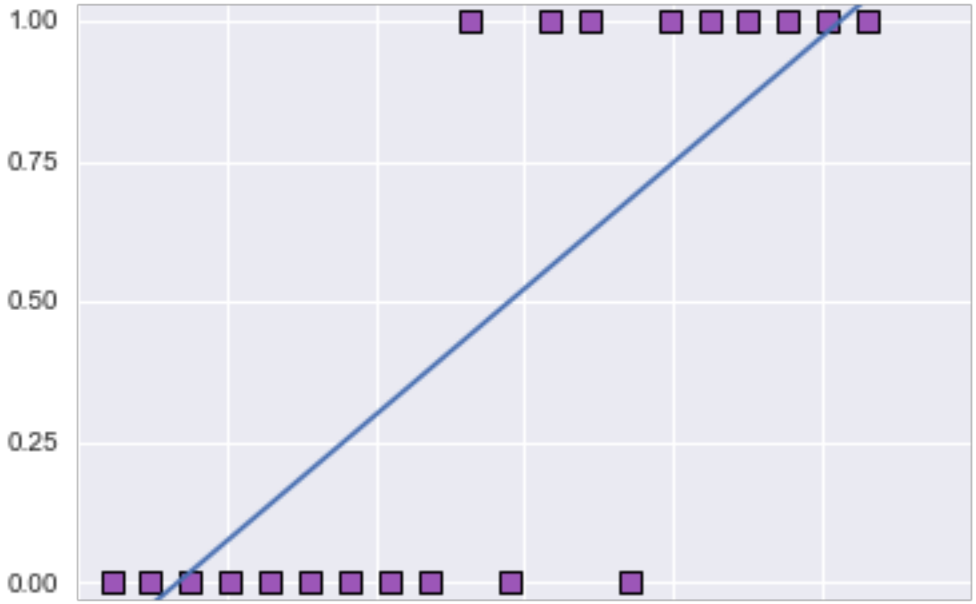
$$\log y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$$



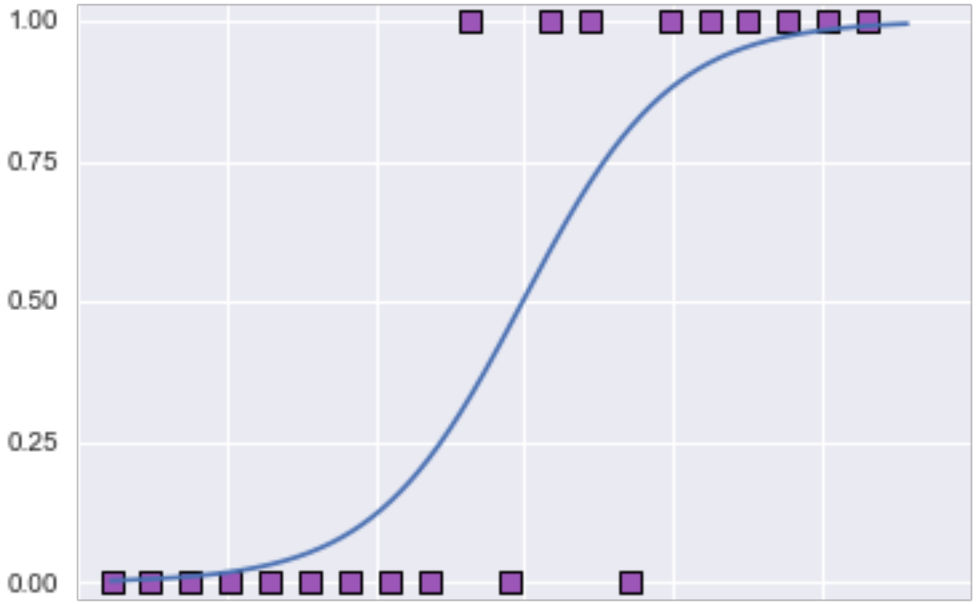
Could we also use regression to classify samples?



Could we also use regression to classify samples?



Could we also use regression to classify samples?



II. LOGISTIC REGRESSION

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Q: What is **logistic regression***?*

Q: What is **logistic regression***?*

A: A generalization of linear regression to classification problems.

In **linear regression***, features predict a continuous outcome variable.*

In **linear regression**, *features predict a continuous outcome variable.*

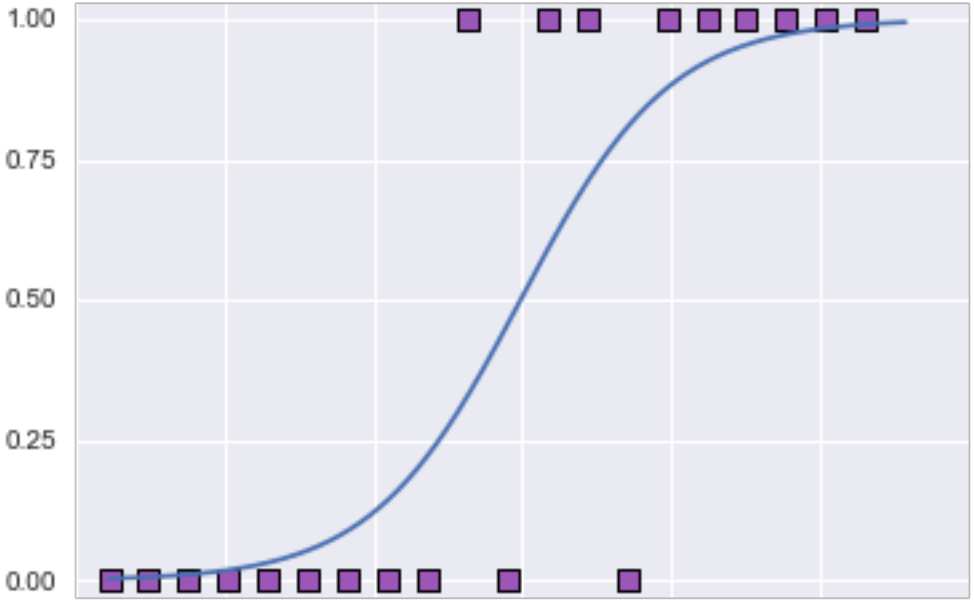
In **logistic regression**, *features predict probabilities of (binary) class membership.*

In **linear regression***, features predict a continuous outcome variable.*

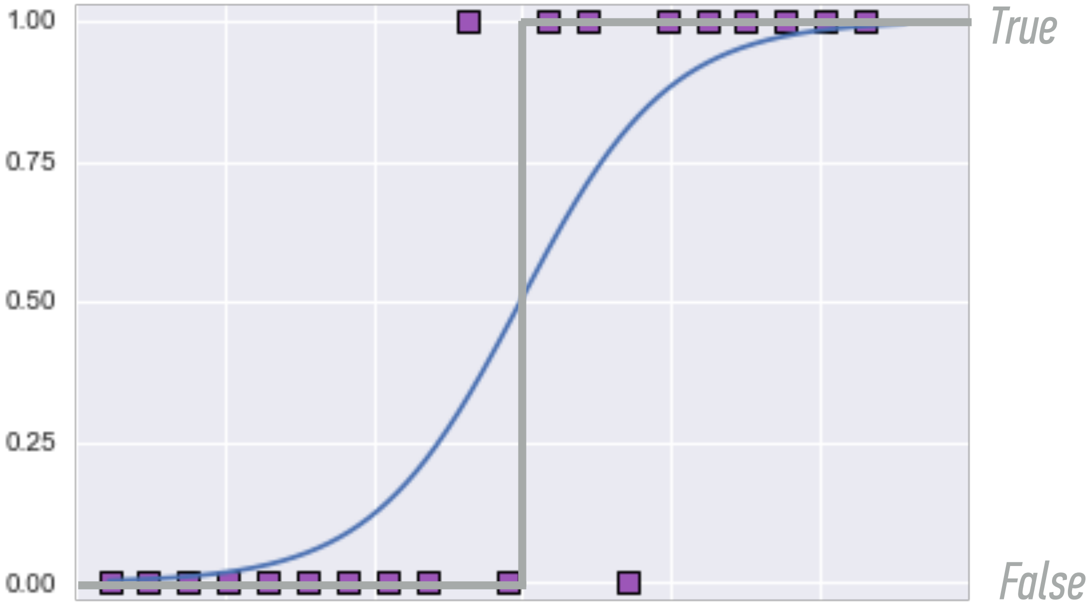
In **logistic regression***, features predict probabilities of (binary) class membership.*

These probabilities are then mapped to class labels, thus solving the classification problem.

Logistic regression gives us predicted probabilities



Logistic regression gives us predicted probabilities, which then could be 'snapped' to class labels



The logistic regression model is an extension of the linear regression model, with a couple of important differences.

The logistic regression model is an extension of the linear regression model, with a couple of important differences.

The first difference is in the outcome variable.

The logistic regression model is an extension of the linear regression model, with a couple of important differences.

The first difference is in the outcome variable.

The second difference is in the error term.

In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval $[0, 1]$.

In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval $[0, 1]$.

So the first step in extending the linear regression model is to map the outcome variable into the unit interval.

In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval $[0, 1]$.

So the first step in extending the linear regression model is to map the outcome variable into the unit interval.

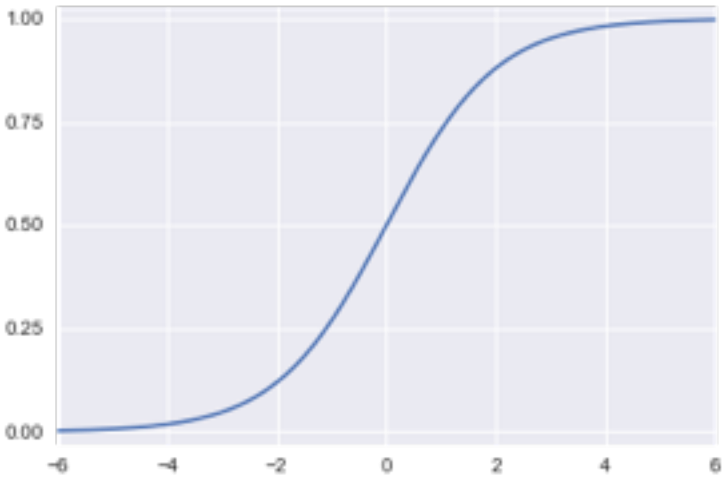
Q: How do we do this?

A: By using a **logistic** *or* **sigmoid function***:*

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

A: By using a **logistic** *or* **sigmoid function***:*

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

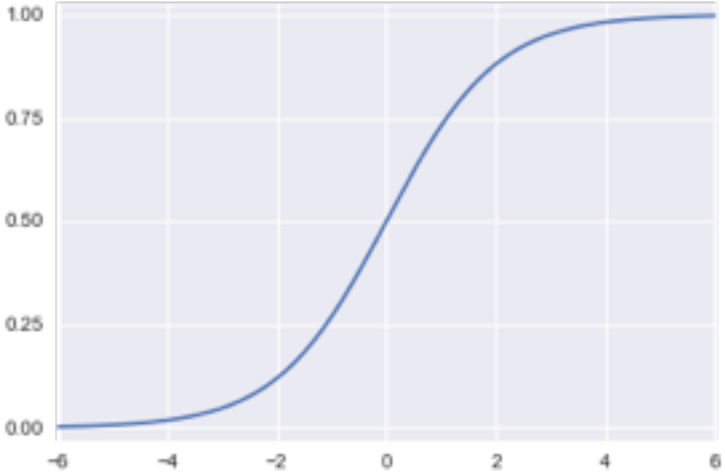


We've already seen what it looks like

A: By using a **logistic** *or* **sigmoid function***:*

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

We've already seen what it looks like

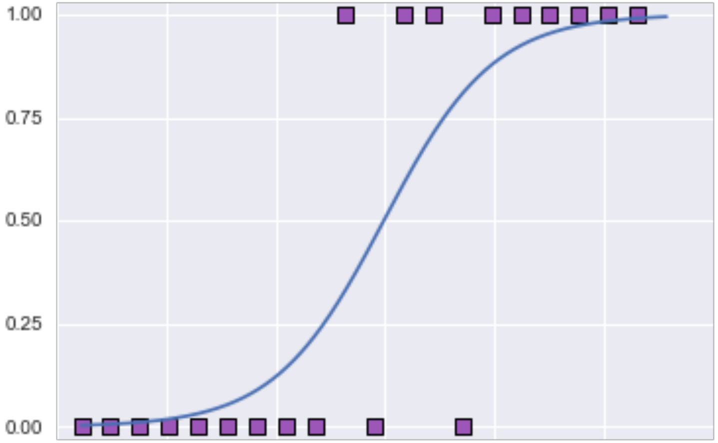


NOTE

For any value of x,
y is in the interval [0, 1]

This is a nonlinear transformation!

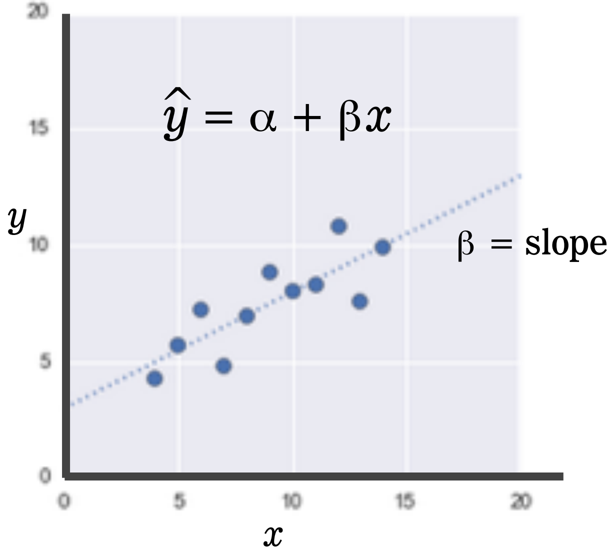
$$y = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \ldots + \beta_n x_n)}} + \varepsilon$$



III. INTERPRETING RESULTS

In **linear regression***, the parameter* β *represents the change in the response variable for a unit change in the* ***X****.*

In **linear regression***, the parameter* β *represents the change in the response variable for a unit change in the* ***X****.*



Example

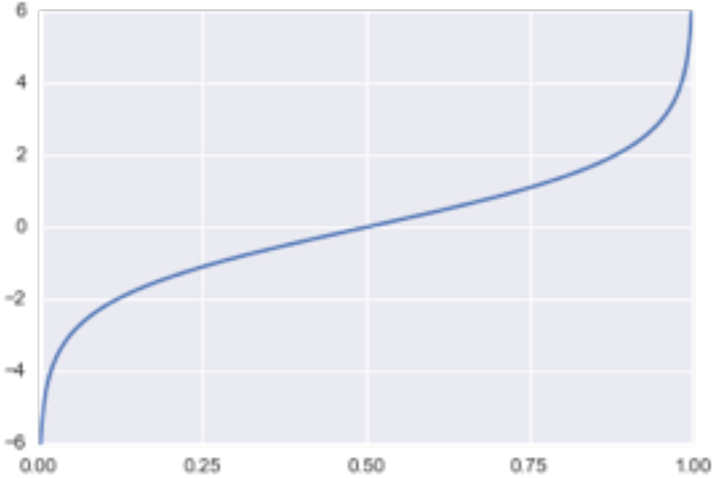
With each additional year work experience, your salary will grow with β *dollars*

In **linear regression***, the parameter β represents the change in the response variable for a unit change in the* ***X****.*

In **logistic regression***, β represents the change in the* **logit** *function for a unit change in the* ***X****.*

The inverse of the logistic function is the **logit** *function,*

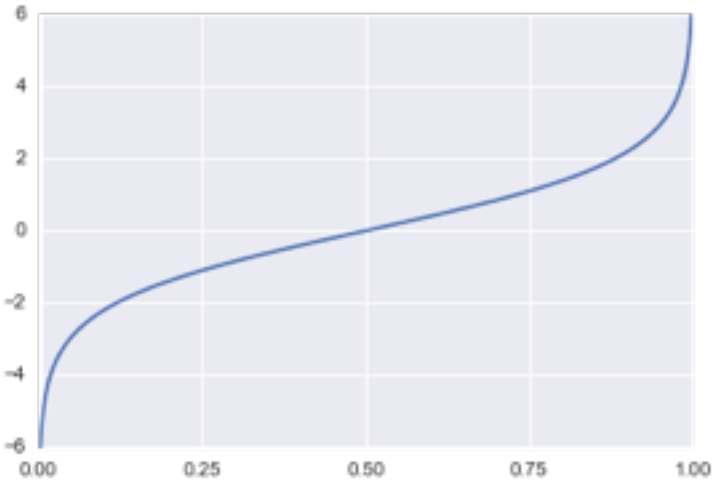
$$\text{logit}(p) = \log \frac{p}{1-p}$$



The inverse of the logistic function is the **logit** *function,*

$$\mathrm{logit}(p) = \log \frac{p}{1-p}$$

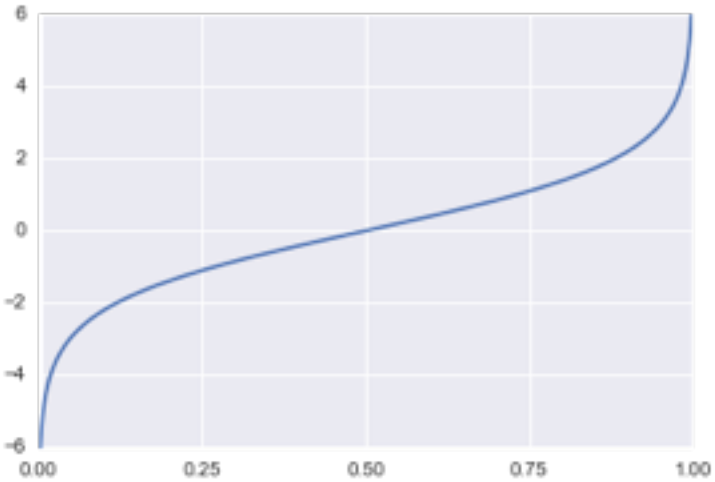
The logit function is also called the **log-odds function**.



The inverse of the logistic function is the **logit** *function,*

$$\mathrm{logit}(p) = \log \frac{p}{1-p}$$

The logit function is also called the **log-odds function.**



NOTE

This name hints at its usefulness in interpreting our results.

We will see why shortly.

In **linear regression***, the parameter* β *represents the change in the response variable for a unit change in the* ***X****.*

In **logistic regression***,* β *represents the change in the logit function for a unit change in the* ***X****.*

Interpreting this change in the logit function requires another definition first.

The **odds** *of an event are given by the ratio of the probability of the event by its complement:*

$$\text{odds}(p) = \frac{p}{1-p}$$

The **odds** *of an event are given by the ratio of the probability of the event by its complement:*

$$\text{odds}(p) = \frac{p}{1-p}$$

The **odds** *tell you how much more likely an event is to happen, compared to the event not happening*

$$\text{odds of } P(A \text{ happens}) = \frac{P(A \text{ happens})}{P(A \text{ doesn't happen})}$$

In **linear regression***, the parameter* β *represents the change in the response variable for a unit change in the* ***X****.*

In **logistic regression***,* β *represents the change in the logit function for a unit change in the* ***X****.*

*In **linear regression**, the parameter β represents the change in the response variable for a unit change in the **X**.*

*In **logistic regression**, β represents the change in the logit function for a unit change in the **X**.*

Example
With each additional year day of training, you will be e^{β} as likely to succeed

In **linear regression***, the parameter* β *represents the change in the response variable for a unit change in the* ***X****.*

In **logistic regression***,* β *represents the change in the logit function for a unit change in the* ***X****.*

Example (for $\beta = \log 2$)
With each additional ~~year~~ day of training, you will be twice as likely to succeed

Suppose we are interested in mobile purchase behavior.

y *= class label denoting purchase/no purchase*

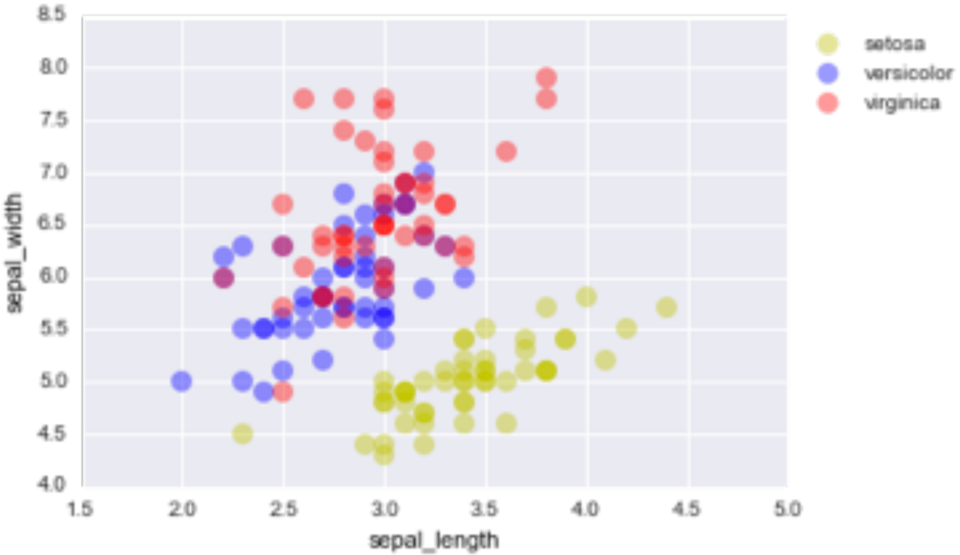
x *= binary flag if a user's mobile OS is Apple's iOS*

Suppose we are interested in mobile purchase behavior.
y *= class label denoting purchase/no purchase*
x *= binary flag if a user's mobile OS is Apple's iOS*

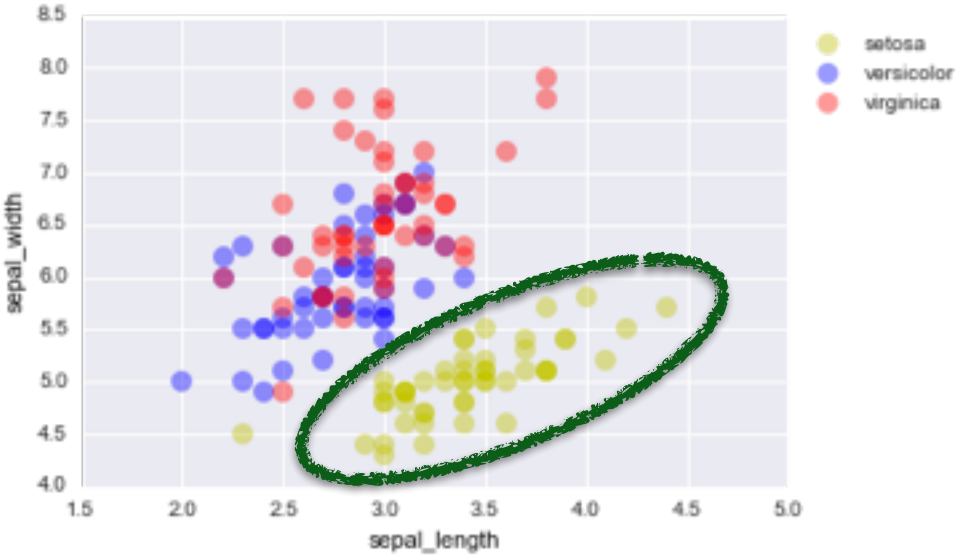
In this case, an odds ratio of 2 (e.g., $\beta = \log 2$*) indicates that a purchase is twice as likely for an iOS user as for a non-iOS user.*

IV. DECISION BOUNDARIES

Let's have a look at the Iris dataset again



Let's predict, using logistic regression, and only sepal's length and width as features, if a flower is a setosa or not

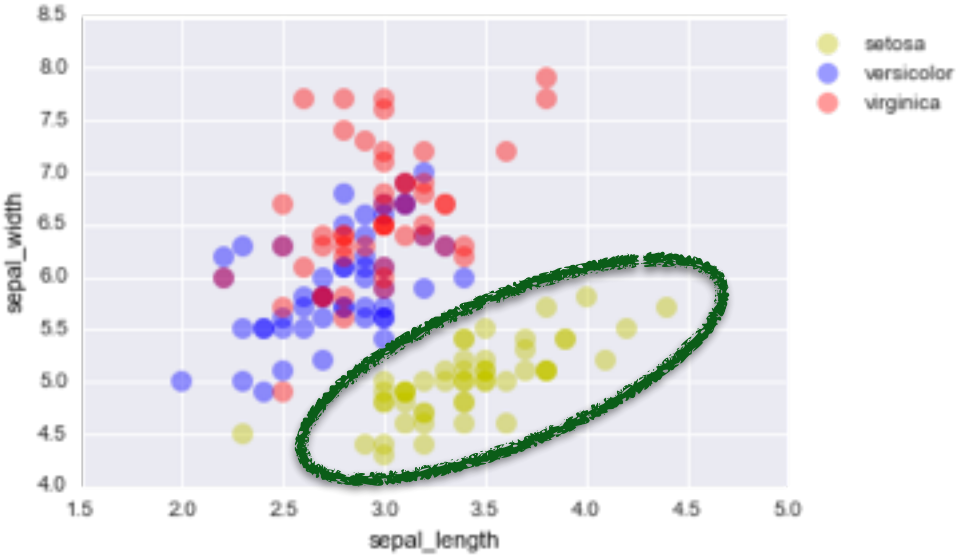


Let's predict, using logistic regression, and only sepal's length and width as features, if a flower is a setosa or not

x_1 = sepal length

x_2 = sepal width

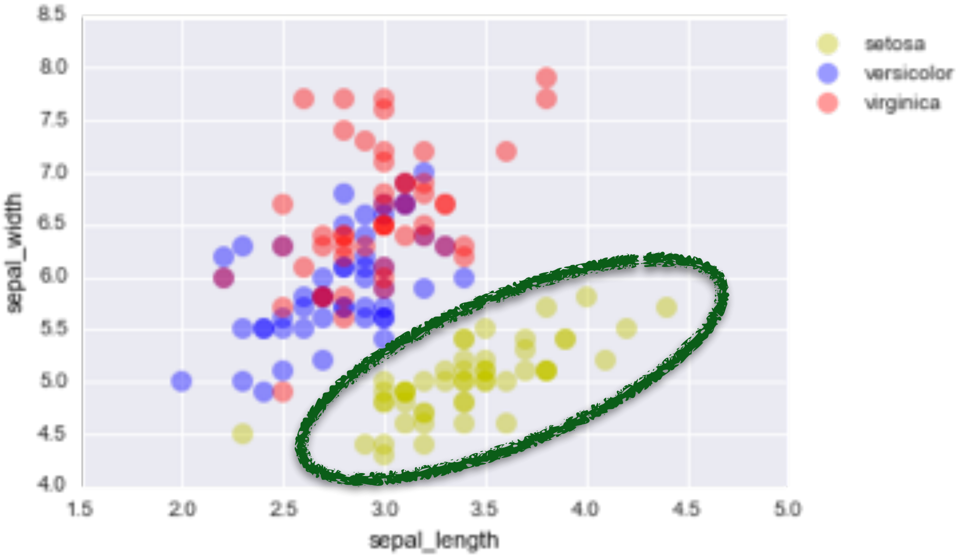
$y = P(setosa)$



$$P(\text{setosa}) = \text{sigmoid}(\alpha + \beta_1 x_1 + \beta_2 x_2)$$

x_1 = *sepal length*

x_2 = *sepal width*



$$P(\text{setosa}) = \text{sigmoid}(\,\alpha + \beta_1 x_1 \;+\; \beta_2 x_2)$$

If we set our cut-off at 50%, we predict a positive iff P ≥ 0.5

$$P(\text{setosa}) = \text{sigmoid}(\,\alpha + \beta_1 x_1 + \beta_2 x_2)$$

If we set our cut-off at 50%, we predict a positive iff $P \geq 0.5$

The case $P = 0.5$ denotes the **boundary** *between our predictions*

$$1/2 \quad = \text{sigmoid}(\alpha + \beta_1 x_1 + \beta_2 x_2)$$

If we set our cut-off at 50%, we predict a positive iff P ≥ 0.5

The case P = 0.5 denotes the **boundary** *between our predictions*

$$0 = \alpha + \beta_1 x_1 + \beta_2 x_2$$

If we set our cut-off at 50%, we predict a positive iff $P \geq 0.5$

The case $P = 0.5$ denotes the **boundary** *between our predictions*

But since $\text{sigmoid}(x) = \dfrac{1}{1 + e^{-x}}$ *it follows that $x = 0$*

In general, for a logistic regression model given by

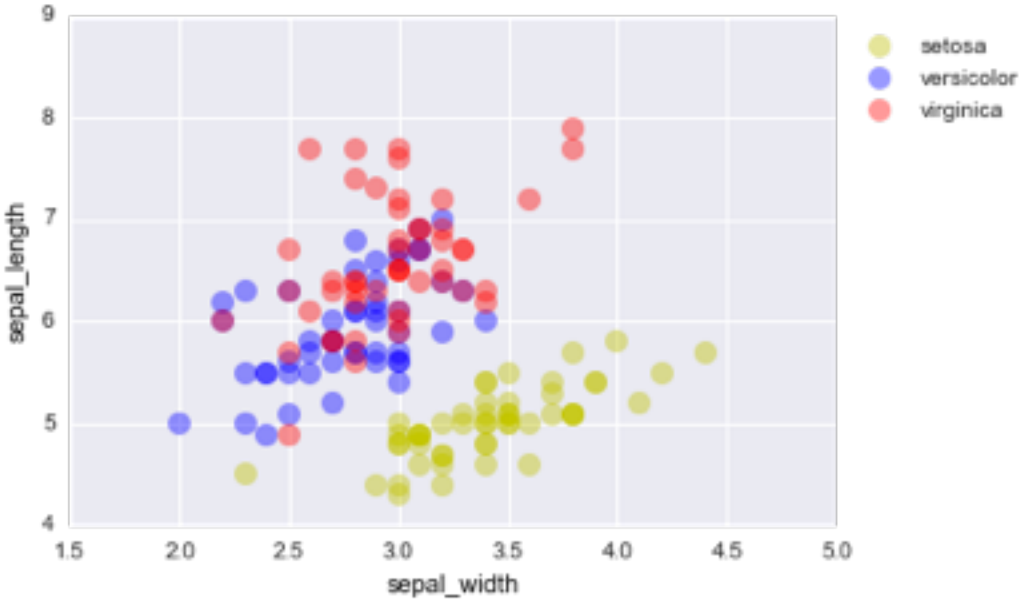
$$\hat{y} = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \ldots + \beta_n x_n)}}$$

its **decision boundary** *is given by the equation*

$$\alpha + \beta_1 x_1 + \ldots + \beta_n x_n = 0$$

$$P(\text{setosa}) = \text{sigmoid}(\alpha + \beta_1 x_1 + \beta_2 x_2)$$

x_1 = *sepal length*
x_2 = *sepal width*
y = P(*setosa*)

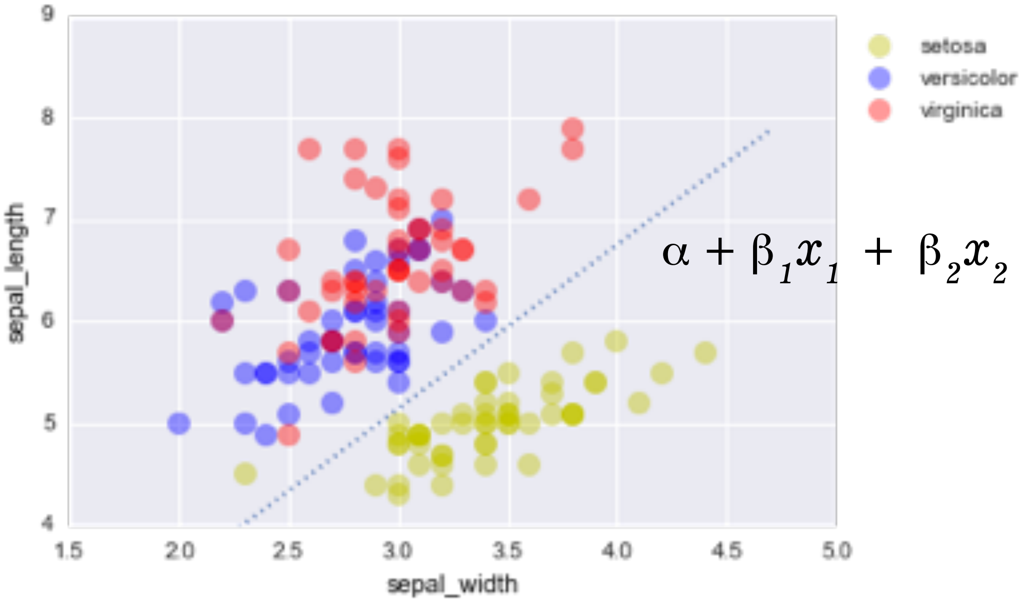


$$P(\text{setosa}) = \text{sigmoid}(\alpha + \beta_1 x_1 + \beta_2 x_2)$$

x_1 = *sepal length*

x_2 = *sepal width*

y = P*(setosa)*



$$\alpha + \beta_1 x_1 + \beta_2 x_2 = 0$$

V. EVALUATING CLASSIFIERS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

Q: How do we evaluate a classification model?

Q: How do we evaluate a classification model?

A. Accuracy

# of times we make the correct classification / # classifications

Q: How do we evaluate a classification model?

A. Accuracy

# of times we make the correct classification / # classifications

When is this a bad a metric?

Q: How do we evaluate a classification model?

A. Accuracy

# of times we make the correct classification / # classifications

When is this a bad a metric?

A: When predicting rare or very likely events

predictions

	Yes	**No**

	predictions	
observations	***Yes***	***No***
Yes		
No		

observations \ *predictions*	**Yes**	**No**
Yes	true positive	
No		

observations	*predictions* **Yes**	**No**
Yes	true positive	
No		true negative

observations \ predictions	Yes	No
Yes	true positive	false negative
No		true negative

observations \ *predictions*	***Yes***	***No***
Yes	true positive	false negative
No	false positive	true negative

$$Accuracy = (TP + TN) / all$$

	predictions	
observ.	Yes	No
Yes	TP	FN
No	FP	TN

CONFUSION MATRIX

Accuracy = *(TP + TN) / all*

Precision = *TP / (TP + FP)*

	predictions	
observ.	*Yes*	*No*
Yes	TP	FN
No	FP	TN

“Of all the cases I highlighted, how often was I right?”

Accuracy = *(TP + TN) / all*

Precision = *TP / (TP + FP)*

Recall = *TP / (TP + FN)*
aka ***hit rate*** or ***sensitivity***

	predictions	
observ.	***Yes***	***No***
Yes	TP	FN
No	FP	TN

“Of all the cases to be highlighted, how many did I hit?”

Accuracy = *(TP + TN) / all*

Precision = *% correct predictions of all positive predictions*

Recall = *% correct predictions of all positive cases*

Accuracy = $(TP + TN) / all$

Precision = % correct predictions of all positive predictions

Recall = % correct predictions of all positive cases

F1 score = $2 \frac{P \times R}{P + R}$

Q: When do you want a high recall model?

Q: When do you want a high recall model?

A. Cost of false positive is low, cost of false negative is high

i.e. Predict who should receive a new cheap drug with low side effects
We want to capture as many people as we can

Q: When do you want a high precision model?

Q: When do you want a high precision model?

A. Cost of false positive is high

i.e. Predict who should receive an expensive, complicated surgery treatment

We want to make sure we are correct on any predictions

DICSUSSION