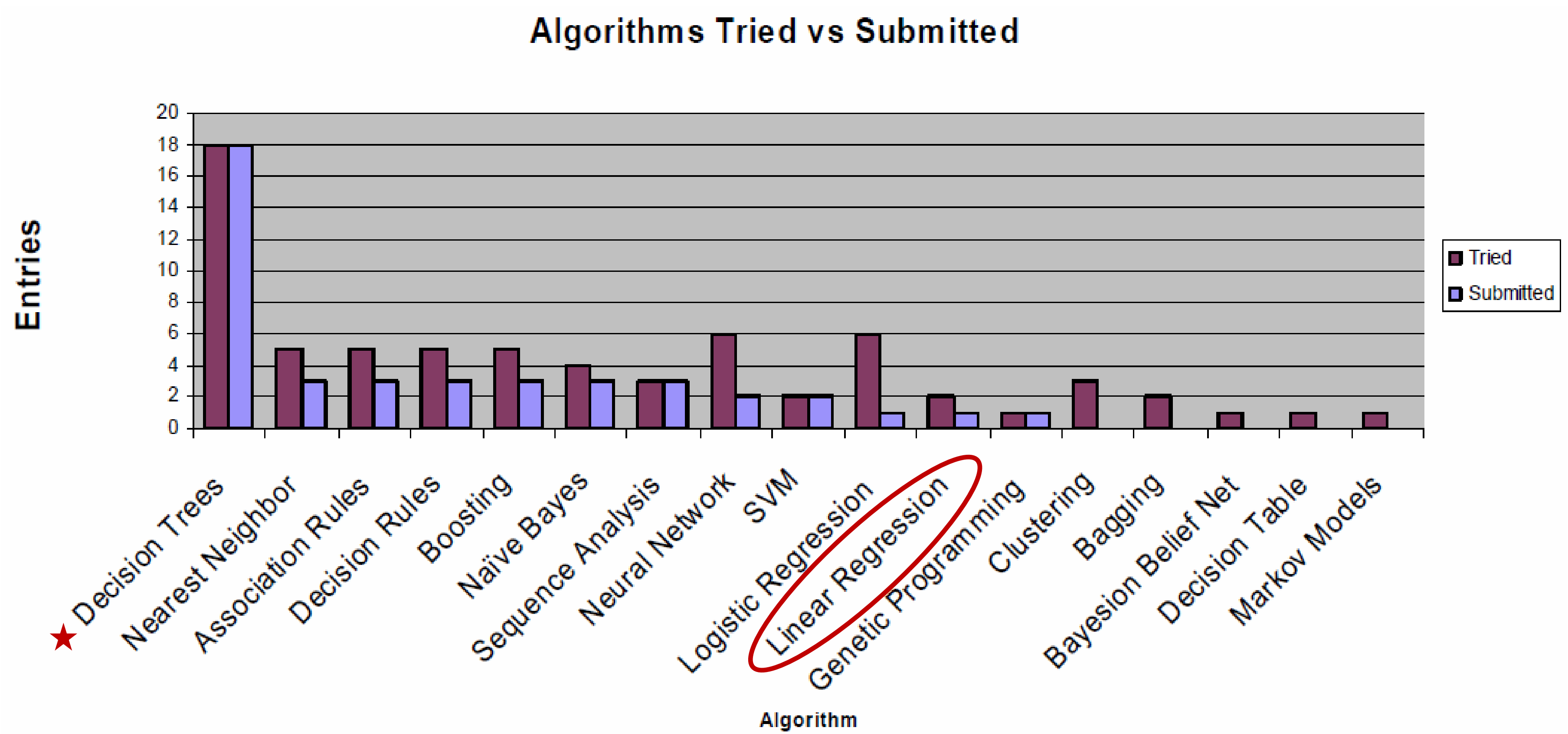
Linear Regression

Instructor: Jing Wang
Department of ISOM

Spring 2023

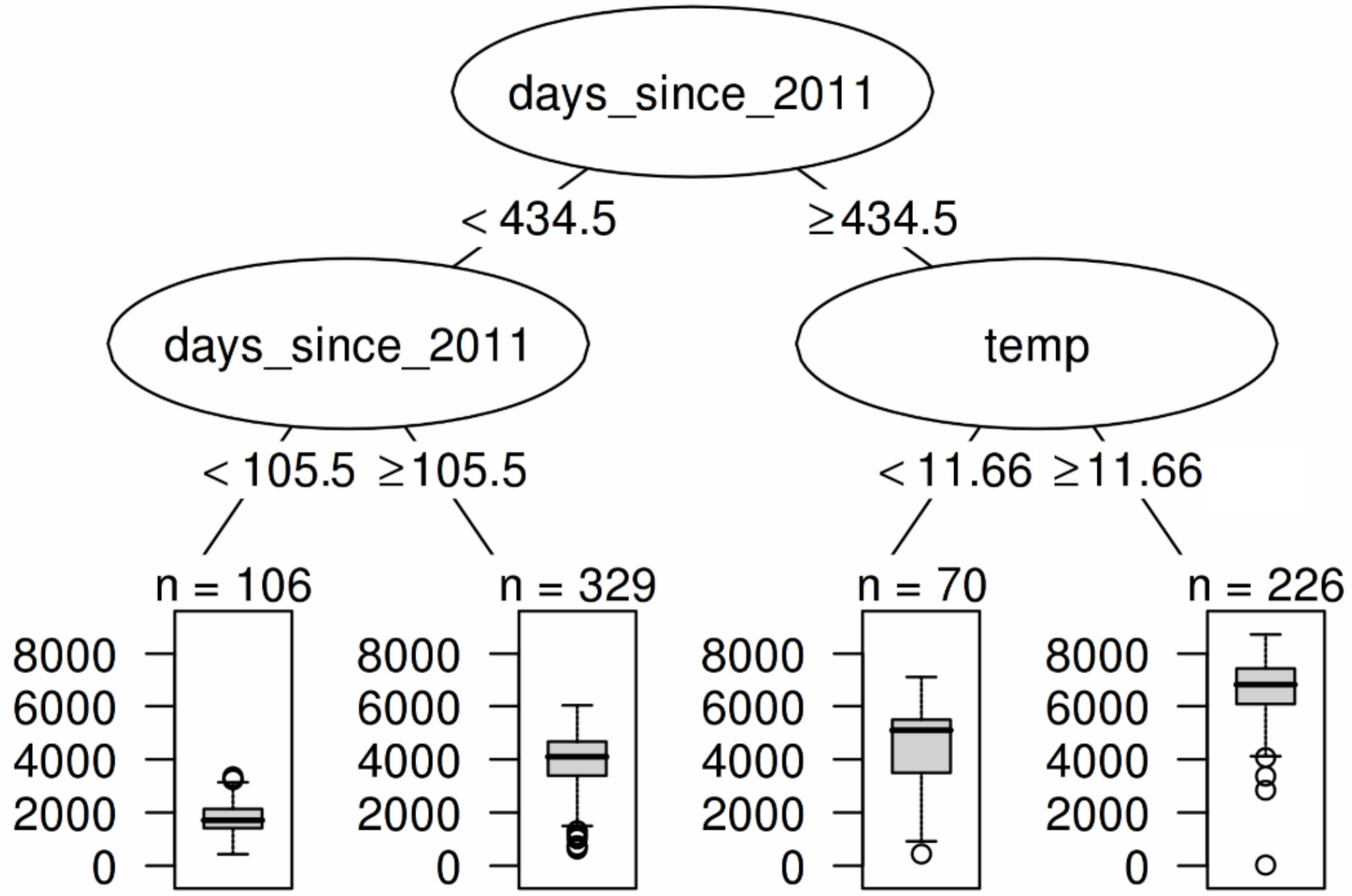
Commonly Used Induction Algorithms



What is Regression?

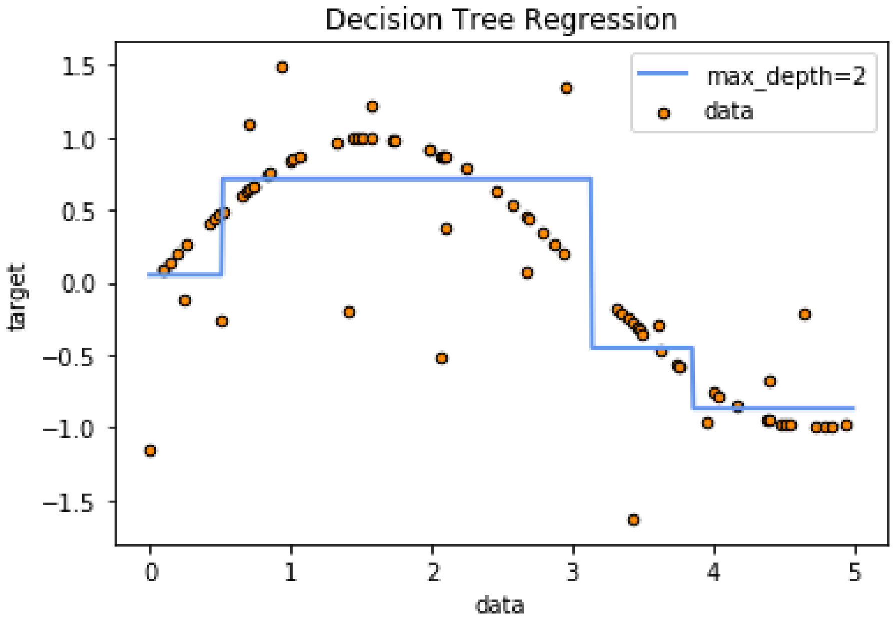
- Regression model in supervised learning predicts a numerical target variable based on a set of predictor variables.
- Applications
  - Predict the impact of discounts on sales in retail outlets.
  - Predict customer credit card activities from their demographics and historical activities.
  - …
- Help businesses make data-driven decisions

Regression Tree (Revisit)

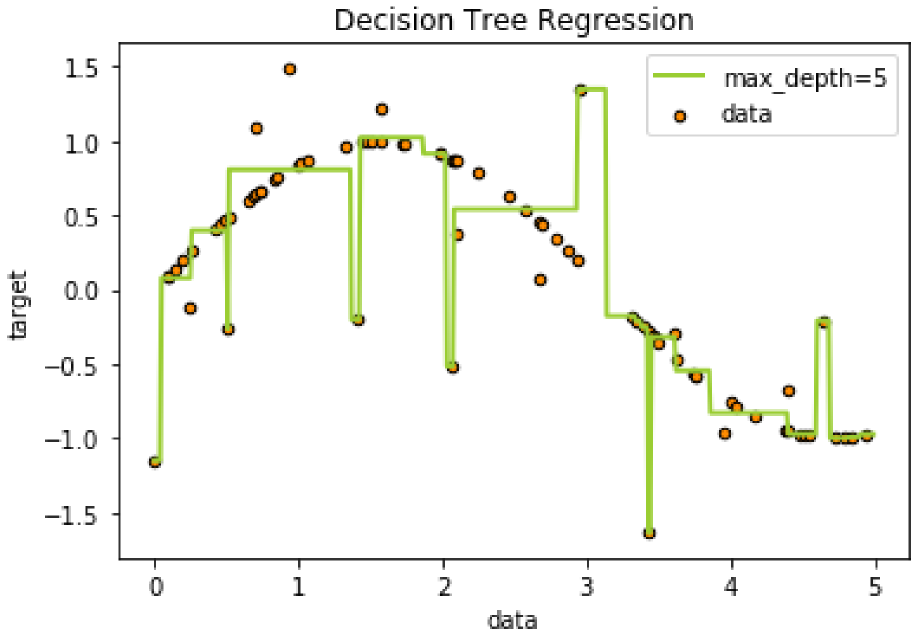


Regression Tree (Reduce Overfitting)

- Stop splitting until max depth is reached.



Tree depth=2



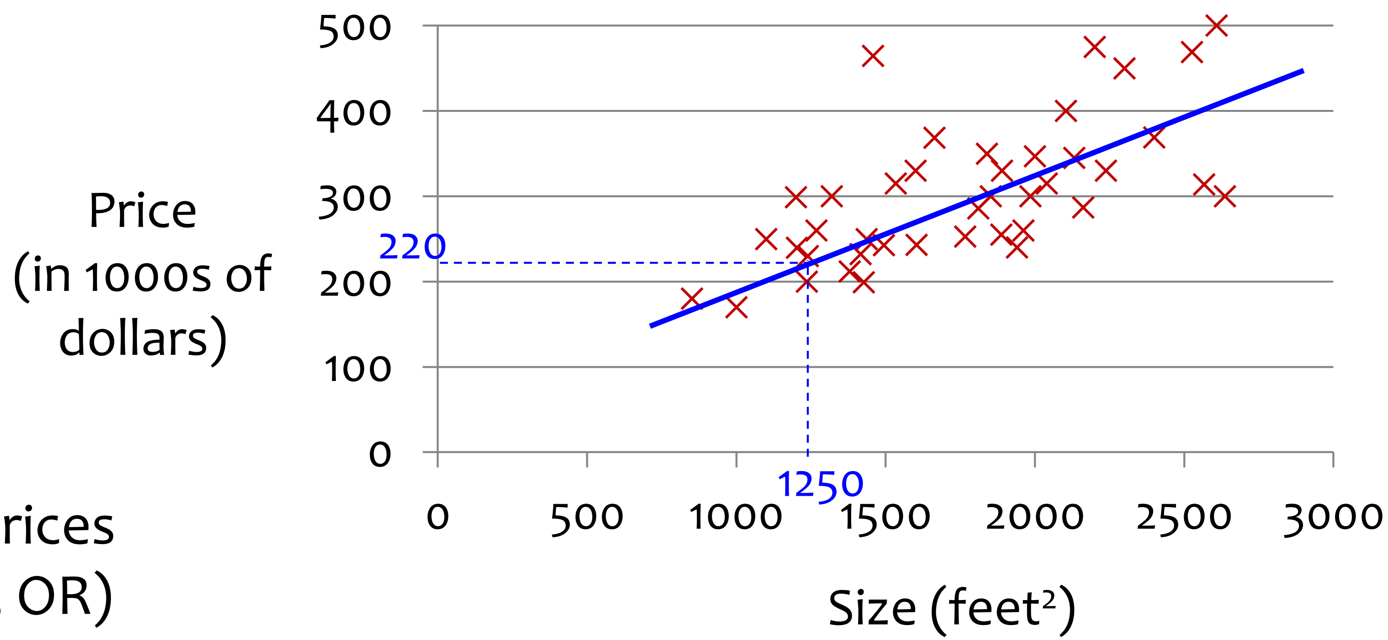
Tree depth=5

Outliers - overfitting

Linear Regression

- Linear regression is the simplest regression model. It models a linear relationship between target variable and predictor variables.

- For example, predict house price based on sq feet.



Housing prices
(Portland, OR)

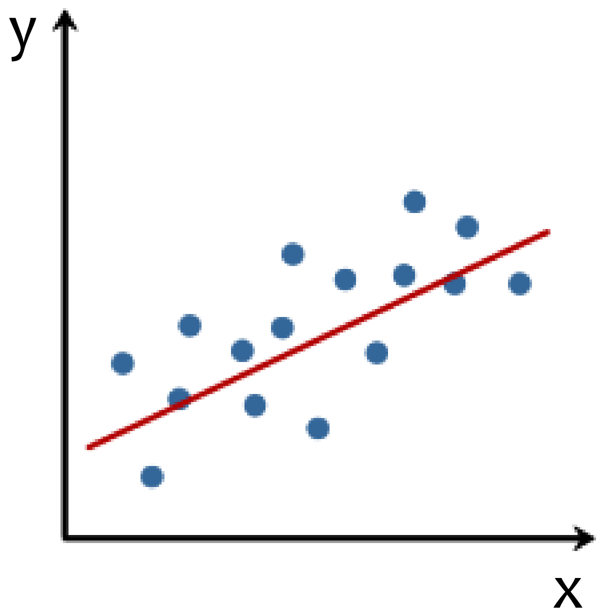
Notations

- Features, attributes, variables: x
- Target variable: y
- Parameters: θ ← Captures the patterns we are looking for
- Predictions: $h_\theta(x)$

x_1 Size (feet²)	x_2 Number of bedrooms	x_3 Number of floors	x_4 Age of home (years)	y Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

Simple Linear Regression

- Simple (univariate) linear regression: only one single predictor variable



$$h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$$

Intercept

Slope

The coefficients we want to know.

The coefficient (except θ_0) represents the change in the target variable for one unit of change in the predictor variable while holding other predictors in the model constant. E.g. Price (in \$1000) = 37.15 + 0.21*Size (in sq ft)

Choose the Right Parameters

$$h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x$$

How to choose θ_i's ?

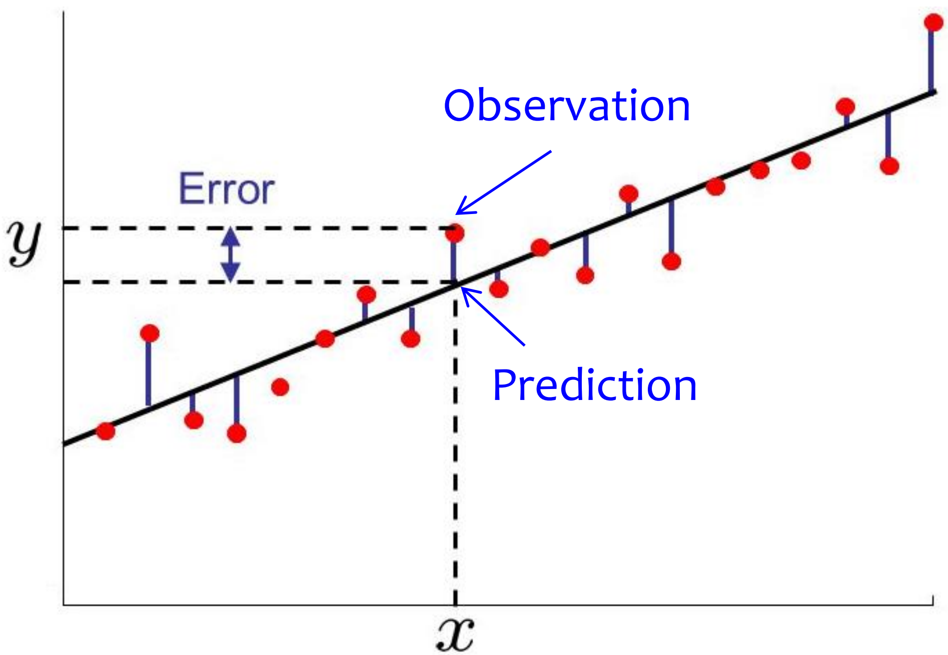
Prediction

Observation

Idea: Choose θ_0, θ_1 so that $h_{\boldsymbol{\theta}}(x)$ is close to y for our training examples (x, y)

Ordinary Least Squares (OLS)

- Ordinary least squares (OLS)
  - Minimizes sum of squared errors



$$\boxed{h_{\boldsymbol{\theta}}(x) = \theta_0 + \theta_1 x}$$

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \boxed{\left(h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)}\right)^2}$$

The squared errors are used to penalize predictions that are far from "true" values.

Multiple Linear Regression

- Multiple linear regression: multiple predictor variables

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

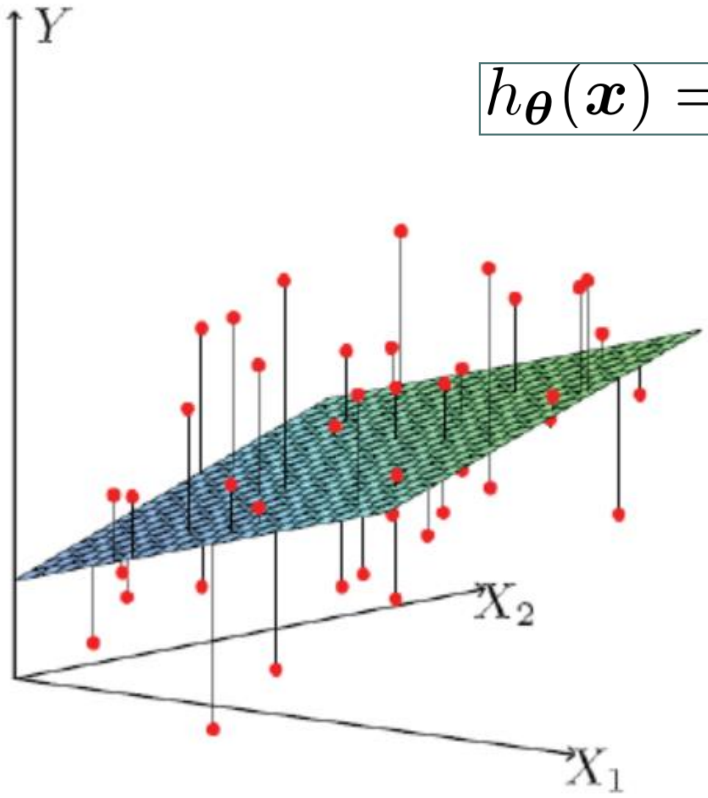
Size (feet²)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

$$\boxed{Price = \theta_0 + \theta_1 Size + \theta_2 Bedrooms + \theta_3 Floors + \theta_4 Age}$$

4 predictor

OLS for Multiple Linear Regression

- Still, ordinary least squares (OLS)
  - Minimizes sum of squared errors



$$\boxed{h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m}$$

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} (h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2$$

The House Price Example

		Coefficient	
x_1	crim	-0.770339	θ_1
x_2	zn	1.048095	θ_2
x_3	indus	0.751485	θ_3
x_4	chas	0.845036	θ_4
x_5	nox	0.641841	θ_5
$\vdots$	rm	3.714814	$\vdots$
	age	-1.535696	
	dis	-1.958528	
	rad	0.060468	
	tax	-1.868513	
	ptratio	-1.702112	
	black	0.678590	
	lstat	-2.065213	

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per 10,000 dollars.
- PTRATIO - pupil-teacher ratio by town
- BLACK - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population

These coefficient values achieve the lowest possible sum of [redacted] squared errors on training examples (the model fitting in linear regression is essentially finding the values of the coefficients)!

Interpreting Coefficients

One unit of change in x_i is associated with θ_i change in the value of y.

- A positive coefficient means that the predictor variable has a positive impact on the value of the target variable, while a negative coefficient means the opposite.
- A large regression coefficient means strong impact on the target variable (note: feature normalization required).

Without feature normalization, does the claim above still hold? Nope！！

Feature Normalization in Linear Regression

- There are a few benefits of applying feature normalization before doing linear regression
  - Ability to rank the importance of features by the relative magnitude of coefficients.
  - A must-do step for regularization (overfitting control).
- Please note:
  - Do the same transformation on your training data and test data!

The Boston House Price Example

	Coefficient
crim	-0.770339
zn	1.048095
indus	0.751485
chas	0.845036
nox	0.641841
rm	3.714814
age	-1.535696
dis	-1.958528
rad	0.060468
tax	-1.868513
ptratio	-1.702112
black	0.678590
lstat	-2.065213

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per 10,000 dollars.
- PTRATIO - pupil-teacher ratio by town
- BLACK - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population

Using absolute value to compare

Which variable has the highest impact on house price? rm

Evaluation Measures for Regression (Recap)

- Mean squared error (MSE)
  - The average of the squares of the differences between predicted values and actual values

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2$$

- Root mean squared error (RMSE)
  - The square root of the average of the squares of the differences between predicted values and actual values

$$\mathbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2}$$

- Mean absolute error (MAE)
  - The average of the differences between predicted values and actual values

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)}|$$

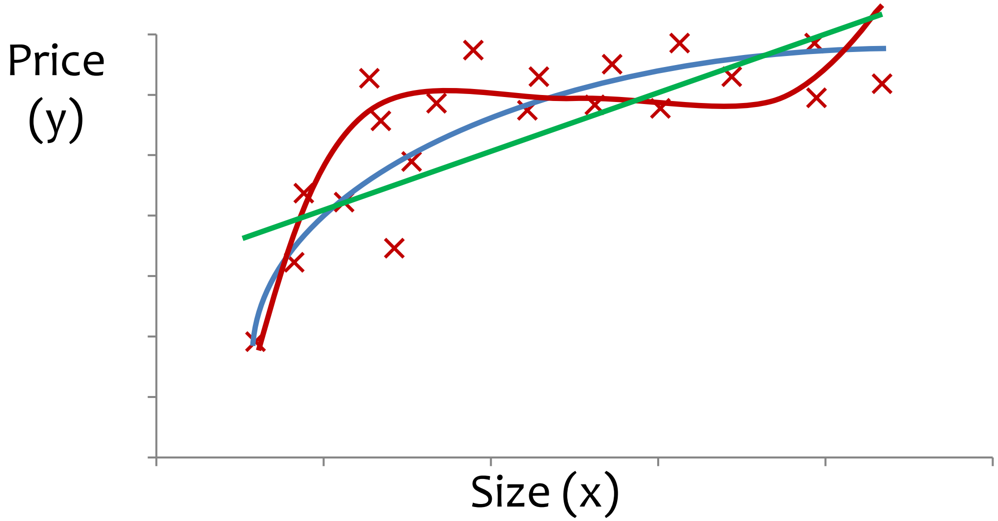
Linear Regression

- Pros
  - Simple to understand and interpret.
- Cons
  - Oversimplifies many real word problems by assuming linearity.
  - Sensitive to outliers.

Practical Issue: More Complex Attributes

- The inputs for linear regression can be
  - Original quantitative inputs
  - Transformation of quantitative inputs
    - e.g., log, square root, square
  - Polynomial transformation
    - e.g., 1, x, x^2, …
  - Interactions between variables
    - e.g., $x_3 = x_1 \cdot x_2$
- "Linear regression" = linear in parameters (θ)

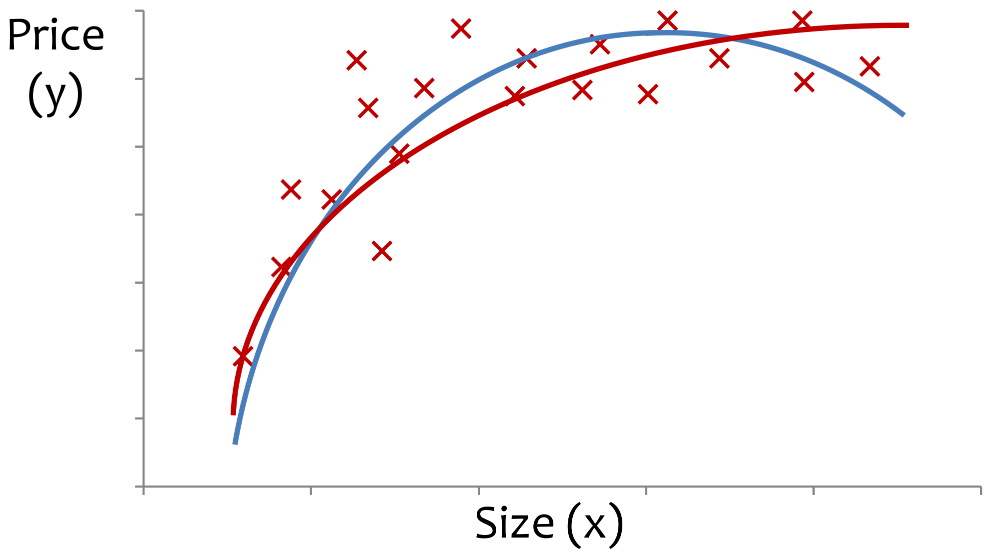
Nonlinearity



$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x$$

blue $$= \theta_0 + \theta_1 x + \theta_2 x^2$$

Red $$= \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$



$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x + \theta_2 x^2$$
$$= \theta_0 + \theta_1 x + \theta_2 \sqrt{x}$$

Using more complex features allows the use of linear regression techniques to fit non-linear datasets.

Practical Issue: Controlling Model Complexity

- Regularization: a method for automatically controlling the model complexity.
  - L1-regularization (Lasso regression)
  - L2-regularization (Ridge regression)
- Idea: penalize for large magnitudes of coefficients
  - Can incorporate into the minimization function
  - Works well when we have a lot of features, each contributing a bit to prediction
- Can address the overfitting problem

Idea Behind Regularization (I)

- All other things being equal, simple models are preferable to complex ones (recall session 5).
  - A simple model that fits the data is unlikely to be a coincidence.
  - A complex hypothesis that fits the data might be a coincidence.

- In linear regression
  - Small values for parameters $\theta_0, \theta_1, \theta_2, \ldots$ = simpler model, which is less prone to overfitting.

Idea Behind Regularization (II)

- Ideally, we want to reduce the magnitudes of coefficients $\theta_0, \theta_1, \theta_2, \dots$ while retaining the model accuracy on the training set.

How can we achieve the two objectives above at the same time?

Answer: extend the minimization function to include the goal of model simplicity (i.e., penalizing large magnitudes of coefficients)

LASSO Regression (L1-Regularization)

number of sample

number of attribute

$$\min_{\boldsymbol{\theta}} \left(\underbrace{\sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)}\right)^2}_{\text{model fit to data}} + \alpha \underbrace{\sum_{j=1}^{m} |\theta_j|}_{\text{L1-regularization}} \right)$$

Linear Regression

Minimize two parts at the same time

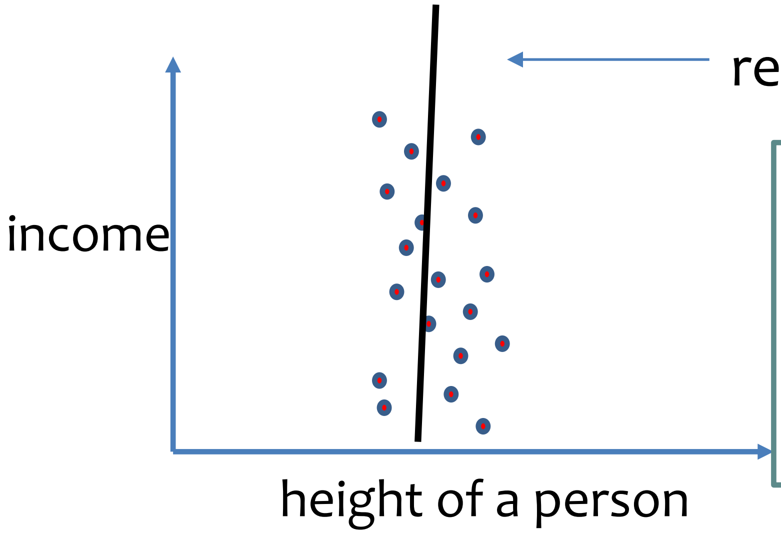
- α is the regularization parameter $(\alpha \geq 0)$
- No regularization on θ_0!

Larger value means stronger penalization.

The estimated coefficients of LASSO regression solve this new minimization function!

What if an attribute does not help reduce the error?

LASSO Regression (L1-Regularization)



regression line with very large coefficient

The height variable has a steep slope (large coefficient), but its impact on income should be very small. The estimated coefficient under LASSO would be small (or even zero).

- LASSO regression penalizes coefficients with large values.
- LASSO regression helps us to pick up the informative attributes (feature selection) by forcing the coefficients of unnecessary attributes to zero.

LASSO Regression -> Feature Selection

- In practice, the dataset could contain hundreds of predictor variables.
- LASSO regression can generate a sparse regression model, which means only a small number of attributes' coefficients are not zero. It addresses overfitting problem by eliminating unnecessary attributes.
- A very rough rule of thumb is to have n>10m, where n is #records, m is #attributes.

LASSO Regression (L1-Regularization)

- Generally, the coefficients of LASSO regression are smaller than the coefficients of linear regression trained on the same set of training examples.
- Because of model simplicity, LASSO regression is less prone to overfitting.
- LASSO regression tends to perform better than linear regression when there are a large number of attributes but not many training examples (e.g., n<=10m, where n is #records, m is #attributes).

Which one is likely to achieve better performance on training examples? Linear regression or LASSO regression?

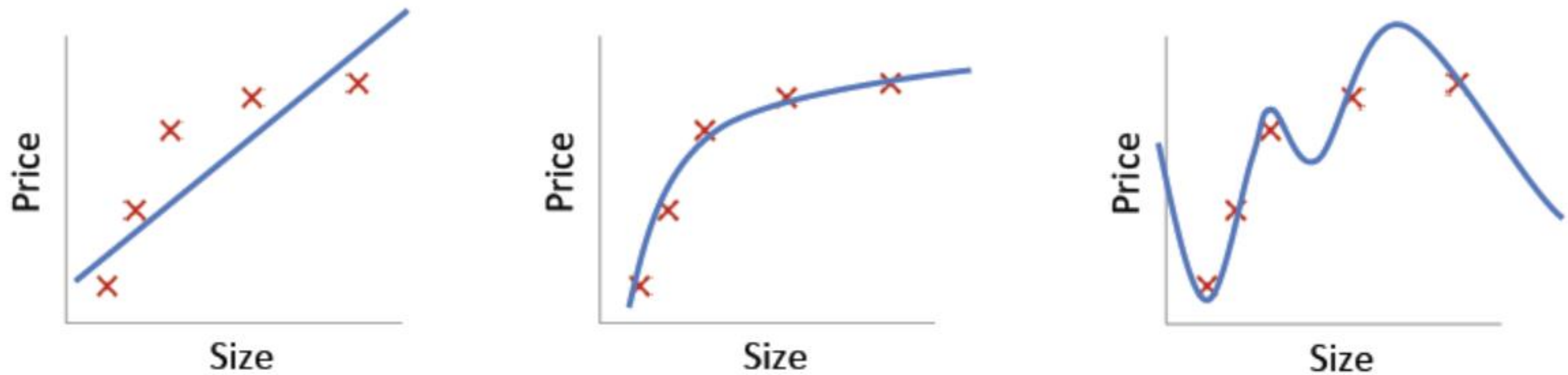
In linear regression only care about the accuracy of model fit

[Optional] Ridge Regression (L2-Regularization)

$$\min_{\boldsymbol{\theta}} \left[\underbrace{\sum_{i=1}^{n} (h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)})^2}_{\text{model fit to data}} + \alpha \underbrace{\sum_{j=1}^{m} \theta_j^2}_{\text{L2-regularization}} \right]$$

- α is the regularization parameter ($\alpha \geq 0$)
- No regularization on θ_0!

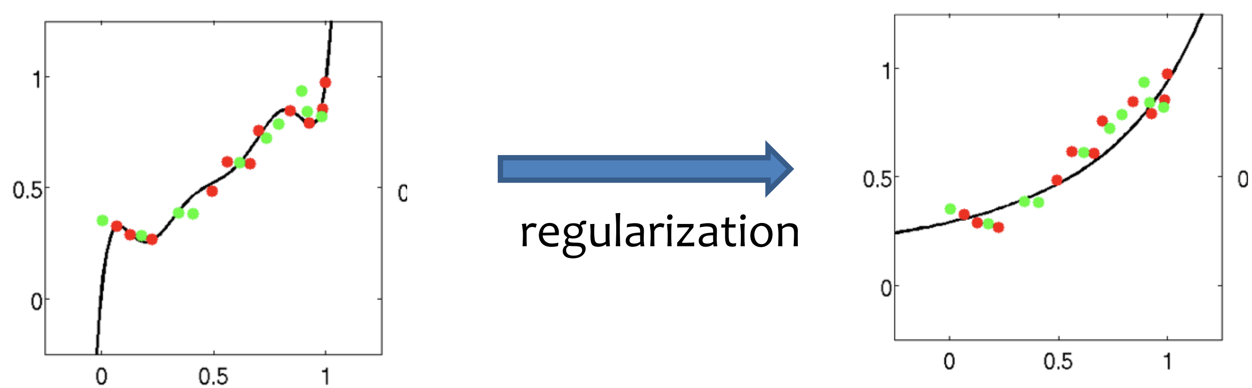
Quality of Model



$\theta_0 + \theta_1 x$ | $\theta_0 + \theta_1 x + \theta_2 x^2$ | $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Underfitting | "Just right" | Overfitting

- The learned model may fit the training set very well (near-zero sum of squared errors), but fails to generalize to new examples.



Lab: Linear and LASSO Regression

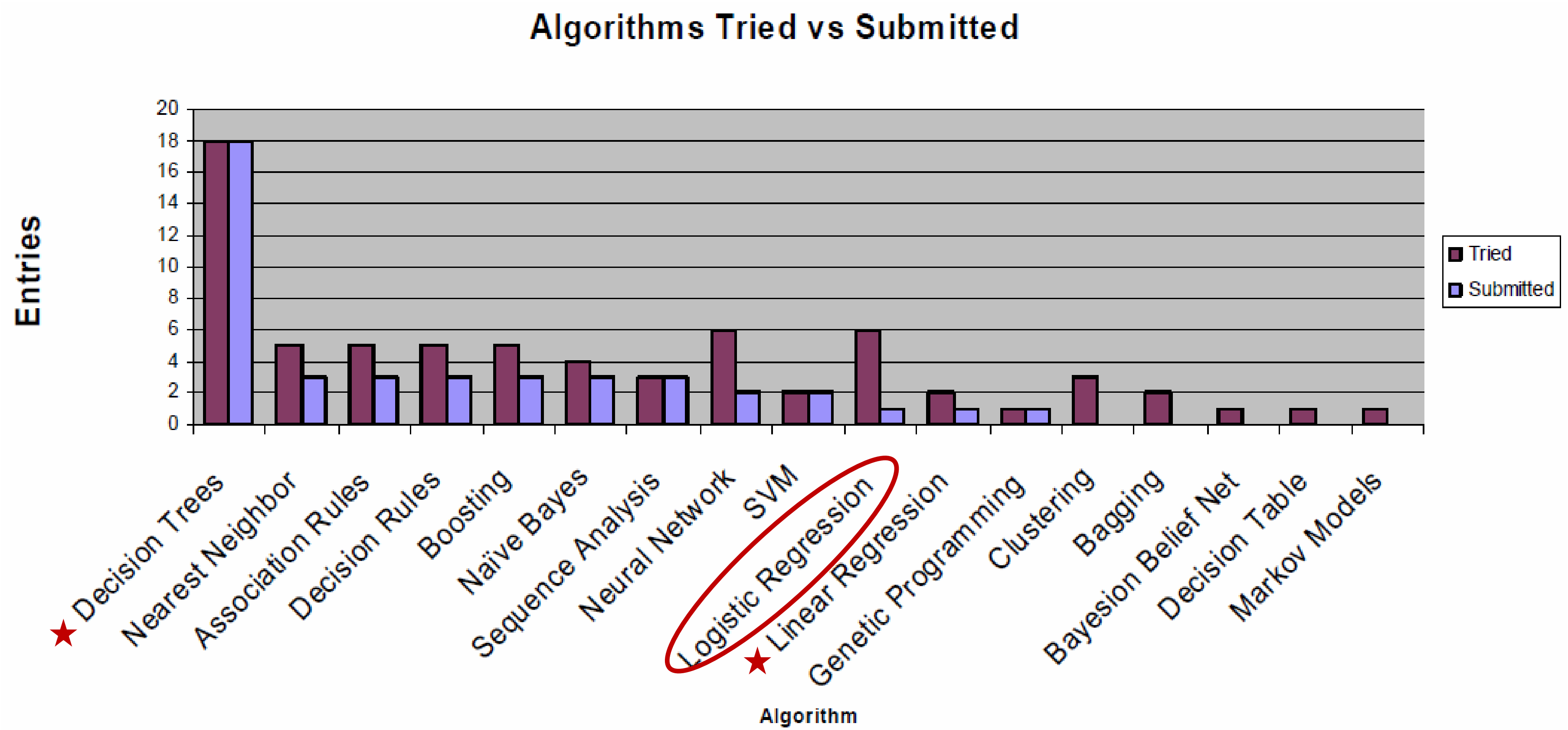
- Files needed
  - LinearRegression.ipynb (Python file)
  - boston.csv (dataset)

Logistic Regression

Instructor: Jing Wang
Department of ISOM
Spring 2023

HKUST BUSINESS SCHOOL 香港科大商學院

Commonly Used Induction Algorithms



Logistic regression is a classification model!

The Term "Regression" in Data Mining

- In data mining, the term "regression" simply means perform prediction on numeric attribute (e.g., house price, spending amount, …)
- It is very different from the meaning of "regression" in statistics. Don't mix them up.
- Decision trees can be used to do regression -> regression trees.
- Not every model with "regression" in name performs regression. Logistic regression is a classification model!

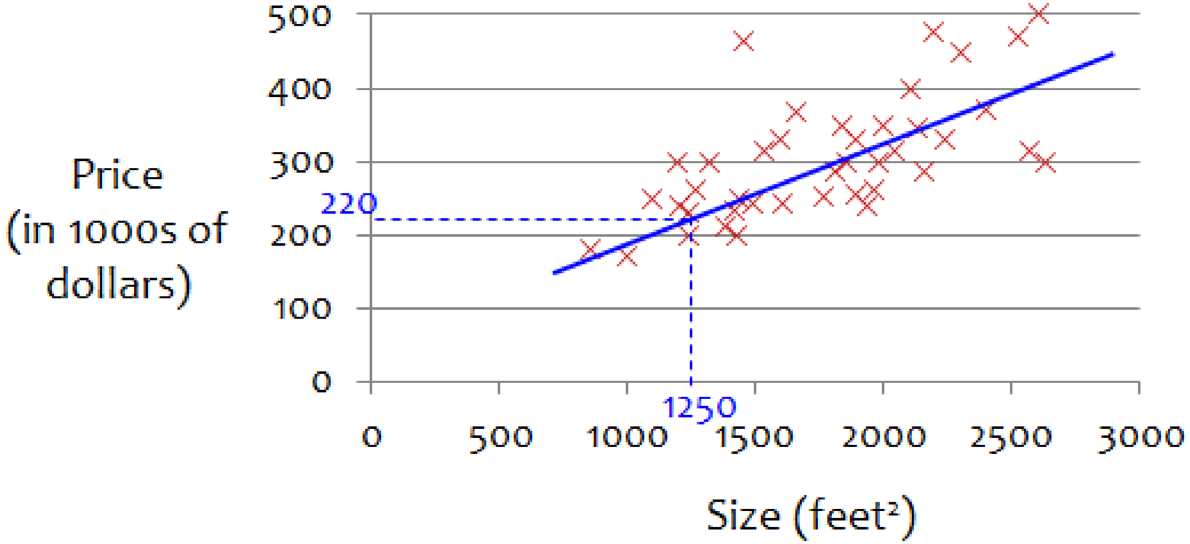
Classification Problems: Revisit

- Churn in cellular services: Stay / Leave?
- Email: Spam / Not Spam?
- Online Transactions: Fraudulent (Yes / No)?

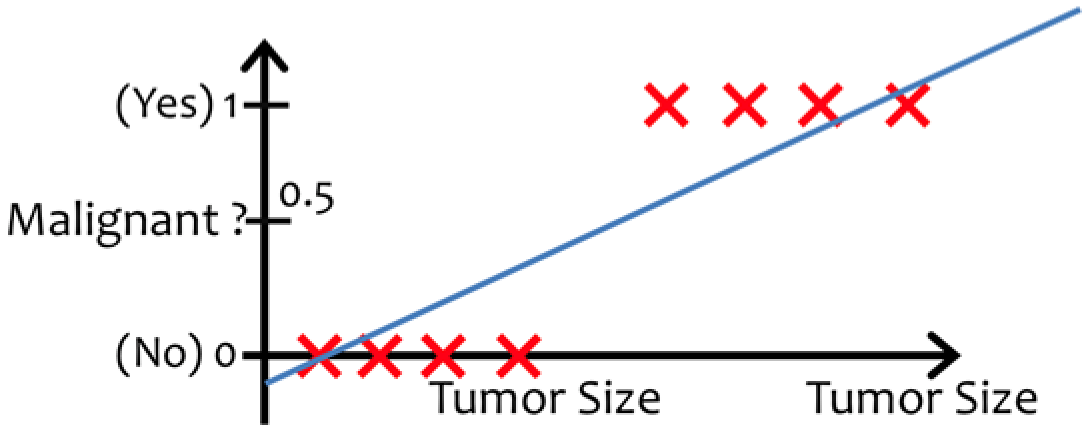
- Tumor: Malignant / Benign ?

$$y \in \{0, 1\} \begin{cases} 0\text{: “Negative Class” (e.g., benign tumor)} \\ 1\text{: “Positive Class” (e.g., malignant tumor)} \end{cases}$$

Linear Regression (Recap)

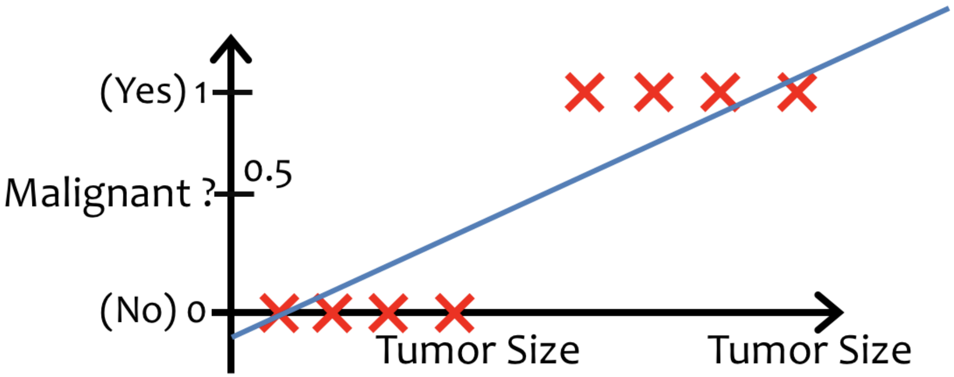


Regression Task: Predict house price on sq feet.



Classification Task: Predict malignant on tumor size.

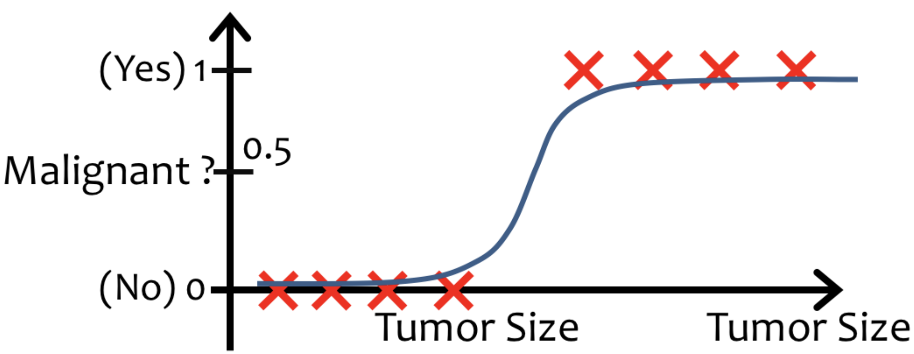
Transformation



$$h_\theta = \theta_0 + \theta_1 TumorSize$$

$h_\theta \in (-\infty, \infty)$

h_θ is monotonically increasing

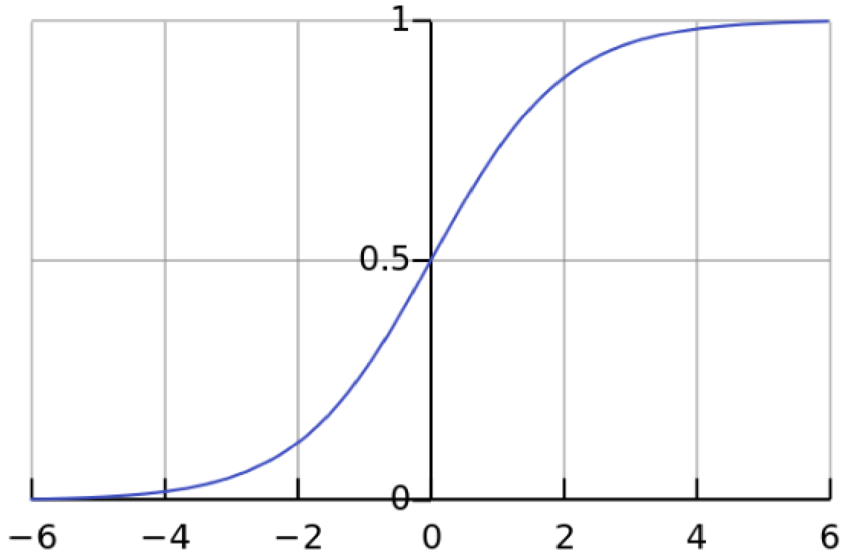


$$h_\theta{}' = g(h_\theta)$$
$$= \boxed{g}(\theta_0 + \theta_1 TumorSize)$$

$h_\theta \in (0,1)$

h_θ is monotonically increasing

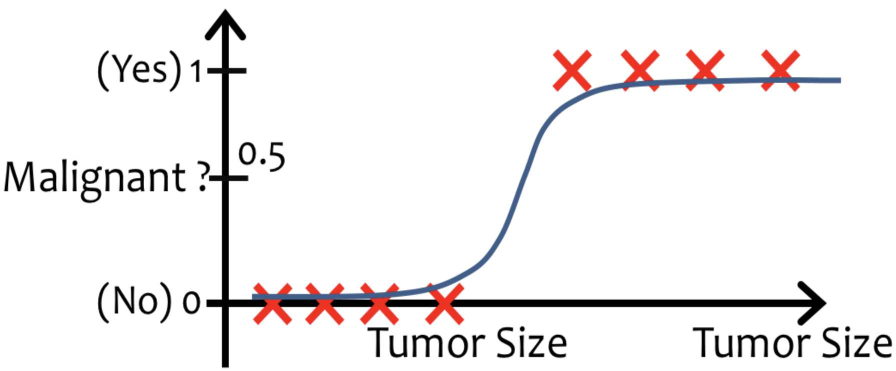
Logistic Function



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

Replace z with the linear regression function

Logistic function, can be between 0 and 1



$$Y(class = 1) = \frac{1}{1+e^{-(\theta_0+\theta_1 TumorSize)}}$$

What does Logistic Regression Do?

- Instead of directly predicting target variable value Y=1 or Y=0, we predict the probability (likelihood) of Y=1, P(Y=1)?

- Given P(Y=1), the probability (likelihood) of Y=0 will be 1- P(Y-1).

- The logistic regression model uses the predictor variables, which can be categorical or continuous, to predict the probability of target variable.

Interpretation of Output

$$h_{\boldsymbol{\theta}}(x) = P(y = 1|x; \boldsymbol{\theta})$$

"probability that y = 1, given x, parameterized by $\boldsymbol{\theta}$"

Example: cancer diagnosis from tumor size

If $h_{\boldsymbol{\theta}}(x) = 0.7$ (x is the size of the tumor),

→ tell patient 70% chance of tumor being malignant.

Note that $P(y = 0|x; \boldsymbol{\theta}) + P(y = 1|x; \boldsymbol{\theta}) = 1$

Therefore $P(y = 0|x; \boldsymbol{\theta}) = 1 - P(y = 1|x; \boldsymbol{\theta})$

Logistic Regression

$$P(y=1|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}$$

- Probability that y=1, given features $x_1, x_2, \ldots, x_m$ parameterized by $\boldsymbol{\theta}$.

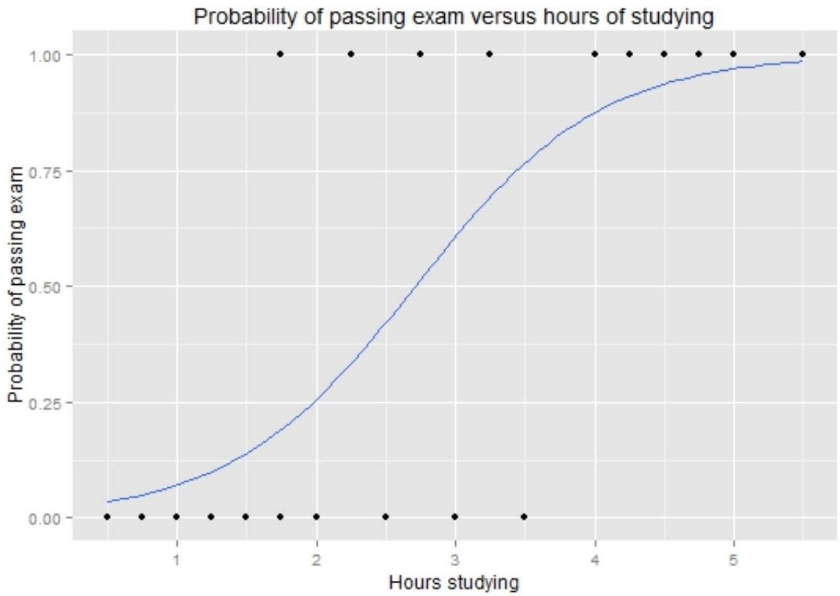
Computer program helps to get the best $\boldsymbol{\theta}$ that "fits" the training data (maximum likelihood estimation).

An Example

- A group of 20 students spend between 0 and 6 hours studying for an exam. Can we predict whether a student will pass an exam based on the hours studying for the exam?

0: failed; 1: passed

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



If a student studies for 2 hours, estimated probability of passing the exam of 0.26;

If a student studies for 4 hours, estimated probability of passing the exam is 0.87.

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot \text{Hours} - 4.0777\right)\right)}$$

A Detour on Odds

- For a given observation, **odds** indicates how much more likely the positive event is to occur than the negative event, e.g., P(Head)/P(Tail) for flipping a coin

Default	Freq.	Percent
Defaulter: 1	49	24.50
Non-Defaulter: 0	151	75.50
Total	200	100

What is the odds of having a Defaulter? 24.5/75.5

Logistic Regression: Another Interpretation

- Logistic regression assumes that log adds of P(Y=1) is a linear combination of coefficients $\boldsymbol{\theta}$ and predictor attributes $\boldsymbol{x}$.

$$P(y=1|\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1+\theta_2 x_2+\cdots+\theta_m x_m)}}$$

$$\Updownarrow$$

$$\log \frac{P(y=1|\boldsymbol{x};\boldsymbol{\theta})}{P(y=0|\boldsymbol{x};\boldsymbol{\theta})} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

Interpreting Coefficients

- Recall: In Linear Regression, how can we interpret the coefficients?

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

- What about coefficients in Logistic Regression?

$$\log \frac{P(y=1|\boldsymbol{x};\boldsymbol{\theta})}{P(y=0|\boldsymbol{x};\boldsymbol{\theta})} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

One unit of change in x_i is associated with θ_i change in the log odds of P(y=1).

Interpreting Coefficients

One unit of change in x_i is associated with θ_i change in the log odds of P(y=1).

- A positive coefficient means that the predictor variable has a positive impact on the probability of the target variable, while a negative coefficient means the opposite.
- A large regression coefficient means strong impact on the probability of target variable (note: feature normalization required).

The Titanic Example

	Coefficient
Pclass	-0.840885
Age	-0.446662
SibSp	-0.369942
Parch	-0.048144
Fare	0.096732
Sex_male	-1.286995
Embarked_Q	0.000000
Embarked_S	-0.196798

Which variable has the highest impact on a person's survival?

Sex

L1 and L2 Regularization in Logistic Regression

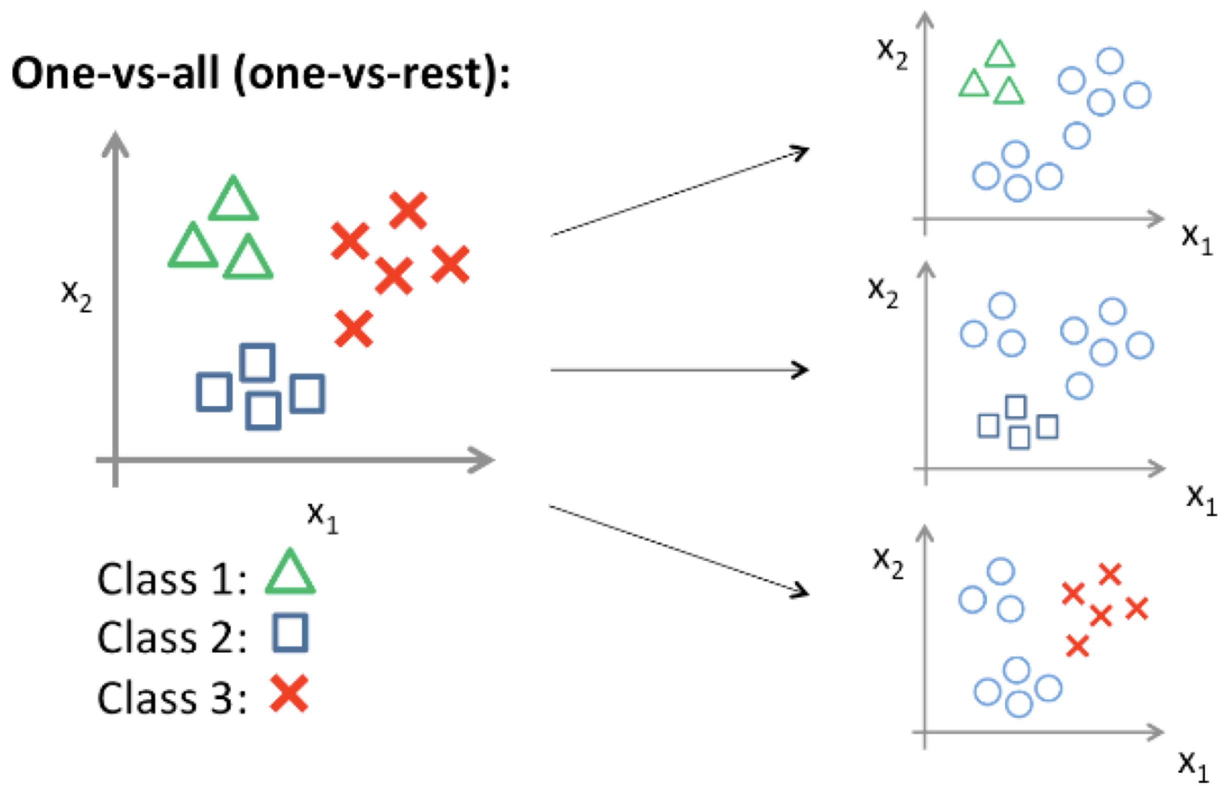
- In logistic regression, we can also use regularization to automatically control the model complexity.
  - L1 regularization (penalize based on absolute values of coefficients)
  - L2 regularization (penalize based on the squared values of coefficients)
- The logistic regression model in sklearn library of Python supports both L1 and L2 regularization (please check the ‘penalty’ parameter when you apply it).

Multi-Class Classification

- So far, we only discuss binary classification, where the target variable only has two values.
- Many of real-world data mining tasks are multi-class classification.
  - Email tagging: work, friends, family, hobby
  - Illness Diagnose: not ill, cold, flu
  - Weather: sunny, cloudy, rain, snow
  - Image Recognition: cat, dog, horse
  - …

Multi-Class Classification

- Solution: For a *k*-class classification, train *k* binary classifiers.
- One-vs-rest (also known as one-vs-all)



$$h_\theta^{(i)}(x) = P(y = i | x; \theta) \qquad (i = 1, 2, 3)$$

One-vs-Rest

- Train a logistic regression classifier $h_\theta^{(i)}(x)$ for each class to predict the probability that $y = i$.

$$h_\theta^{(i)}(x) = P(y = i|x;\theta)$$

- On a new example x, to make a prediction, pick the class i that maximizes

$$\max_i h_\theta^{(i)}(x)$$

$P(y=1|x;\theta)$ = 0.6

$P(y=2|x;\theta)$ = 0.4

$P(y=3|x;\theta)$ = 0.2

What is your prediction?

Lab: Logistic Regression

- Files needed
  - LogisticRegression.ipynb (Python file)
  - titanic_cleaned.csv (dataset)

Generalization Performance

- Different modeling procedures may have different performance on the same data.
- Different training sets may result in different generalization performance.
- Different test sets may result in different estimates of the generalization performance.
- If the training set size changes, you may also expect different generalization performance from the resultant model.

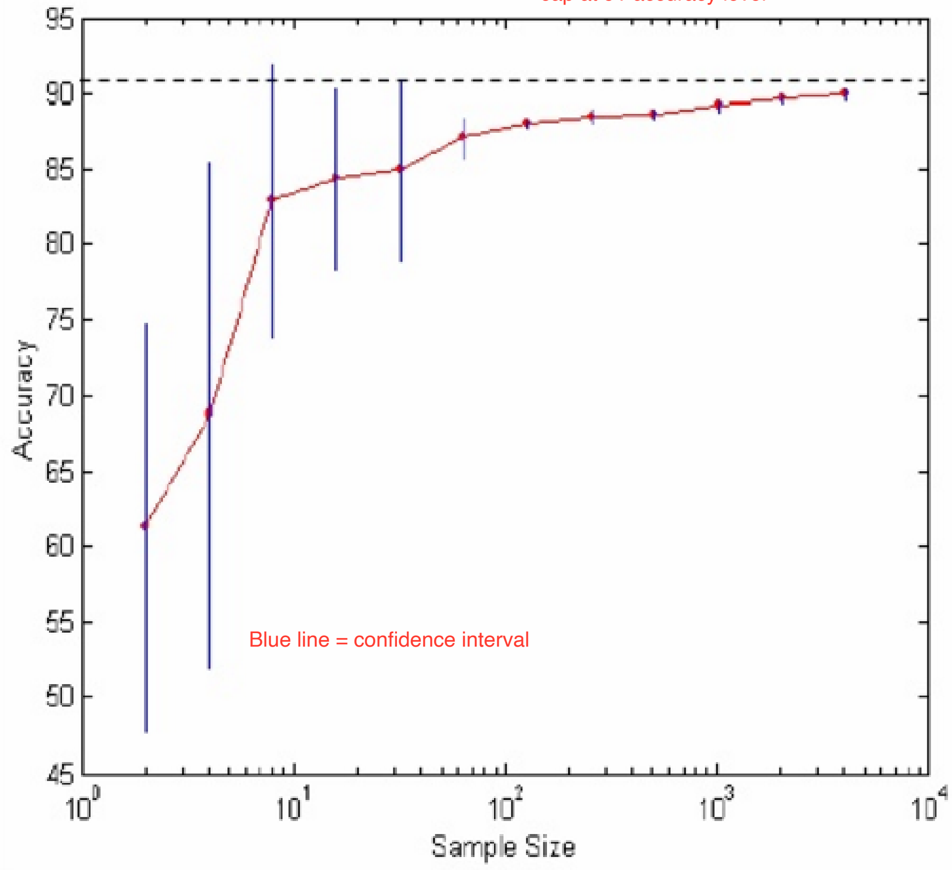
Learning Curve

A learning curve shows how the generalization performance changes with varying sample size!

More example, better performance (accuracy increase)

When sample size large enough, keep increasing the sample size will not affect the accuracy a lot

cap at 91 accuracy level



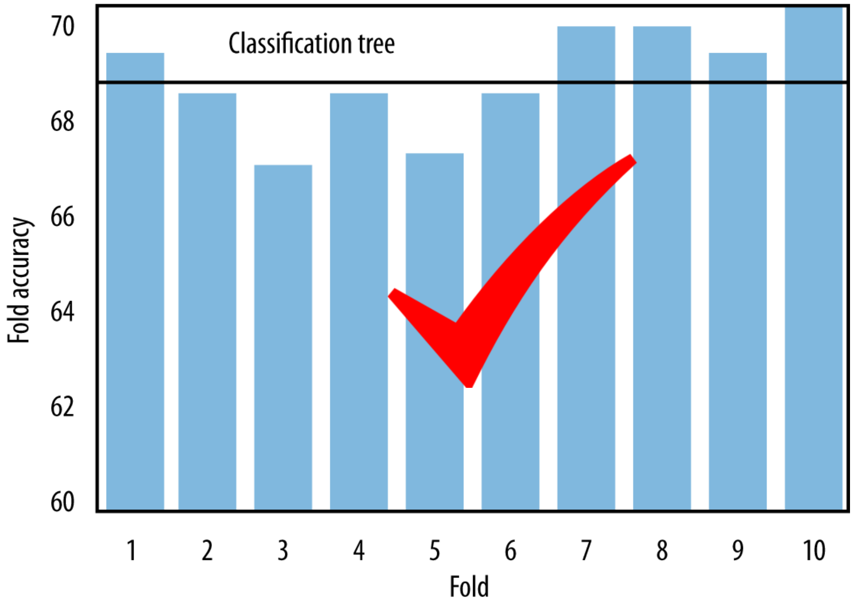
Decision Trees vs. Logistic Regression (I)

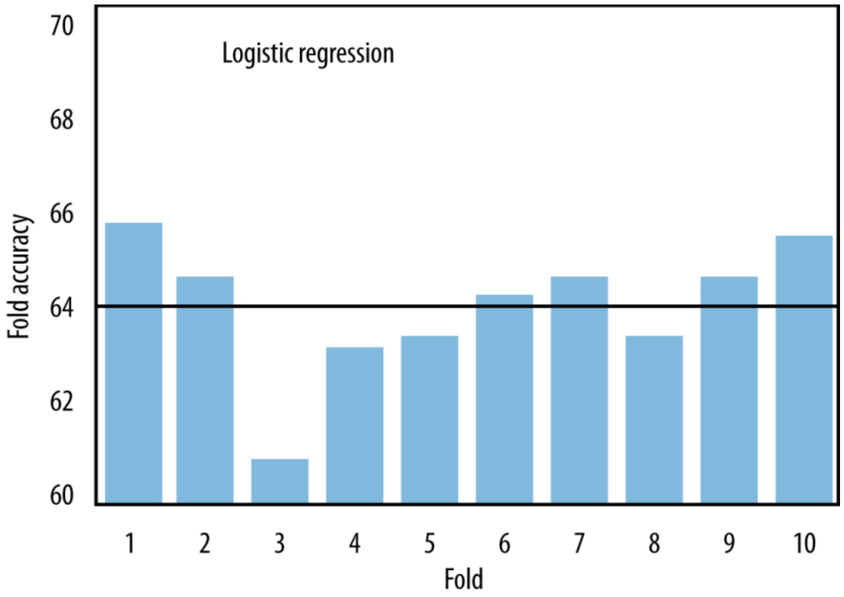
- What is more comprehensible to the stakeholders?
  - Rules?
  - Numeric functions?
- How much data do you have?
  - For smaller training-set sizes, logistic regression yields better generalization accuracy than tree induction
  - With larger training sets, flexibility of tree induction can be an advantage: trees can represent substantially nonlinear relationships between the features and the target (need pruning to reduce overfitting)

A Real Case: TelCo

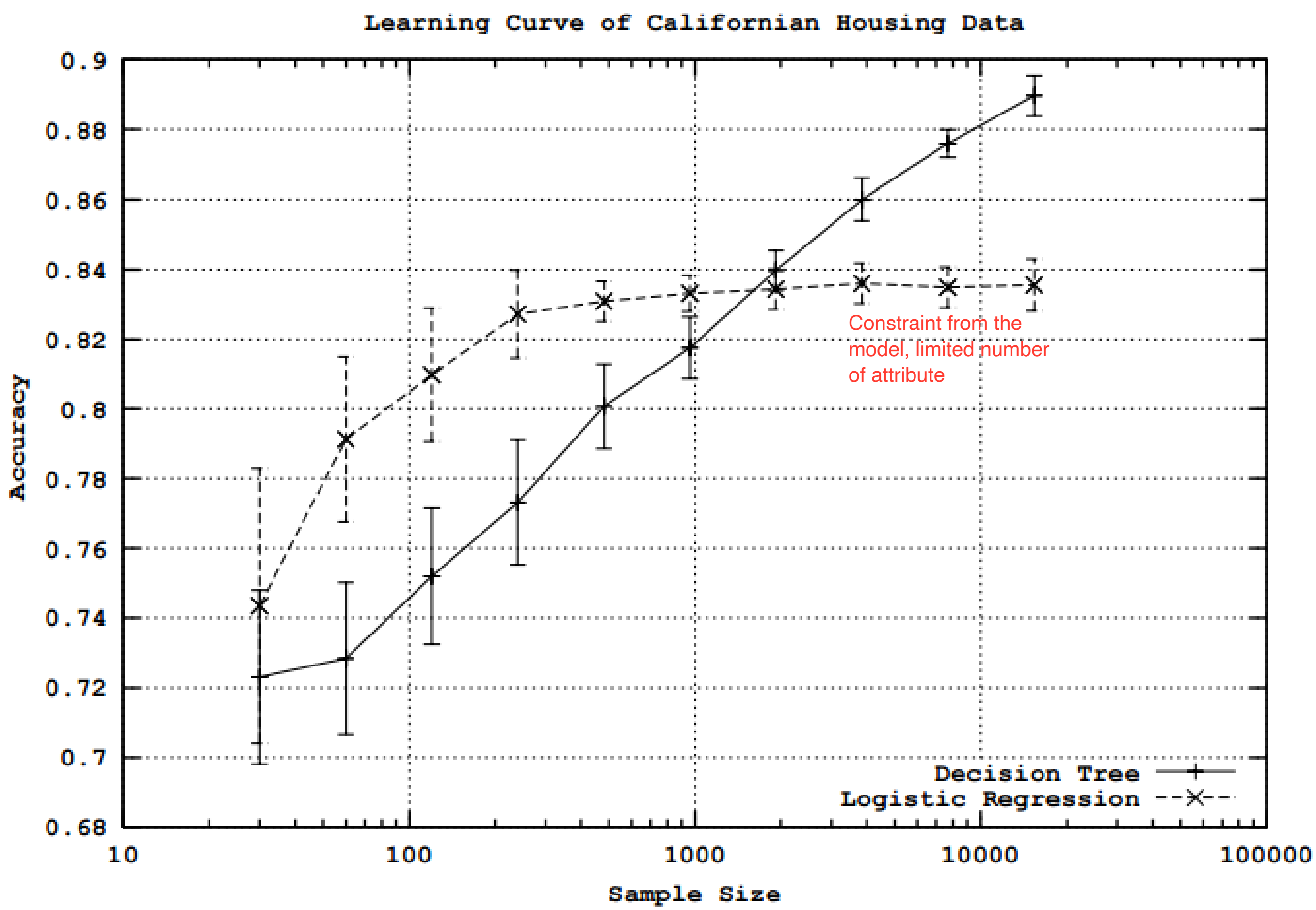
- TelCo, a major telecommunications firm, wants to investigate its problem with customer attrition, or “churn”. They want to build a model to predict the churning probability of customers.

This dataset contains 20,000 examples.





Learning Curve Comparison



Decision Trees vs. Logistic Regression (II)

- What are the characteristics of the data?
  - Trees are fairly robust to: missing values, types of variables (numeric, categorical), how many are irrelevant, etc.
  - Trees do not perform well when there is a lot of noise in the data.
- Do you need good estimate on class probabilities?
  - Logistic regression generates probabilities in a more sophisticated way.

Do you still remember how tree generates class probability estimates?