AGENDA

- Homework and project
  - Due Saturday
  - Talk to me
- Answer exit survey tickets
- Review Last time (ML, regression)
- Continue with Linear regression
  - Multiple linear regression
  - Interactions
  - Categorical
  - Non-linear
- Training and testing
- KNN

QUESTIONS FROM EXIT SURVEY

- P-Values
- What statements do we need to make concerning the accuracy of prediction obtained from our model
- How does linear regression compare to other models?
  - Speed?
  - Polynomial, non-linear regression?
- Assumptions of linear regression
  - Checking assumptions

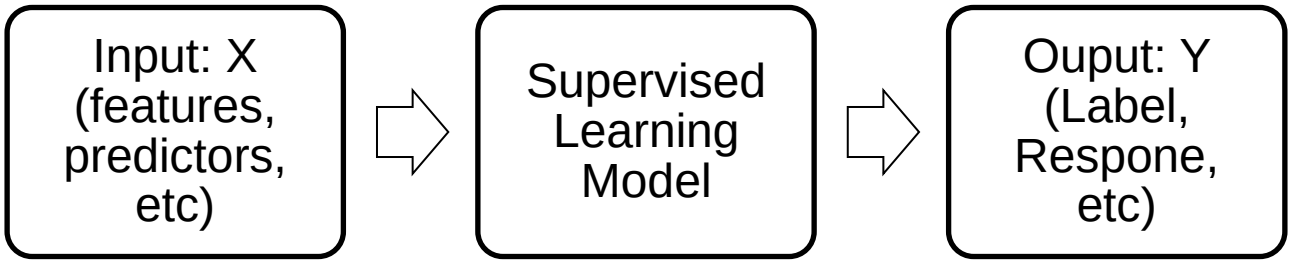
REVIEW FROM LAST TIME

- Intro to Machine Learning
  - What is machine learning
  - Different types of machine learning
  - Parts of machine learning
  - Machine learning in practice
- Vocabulary
  - Observations
  - Features, predictors, attributes
  - Response, labels
  - Dimensionality
  - Supervised, unsupervised, regression, classification
- Simple linear regression
- Multivariate linear regression

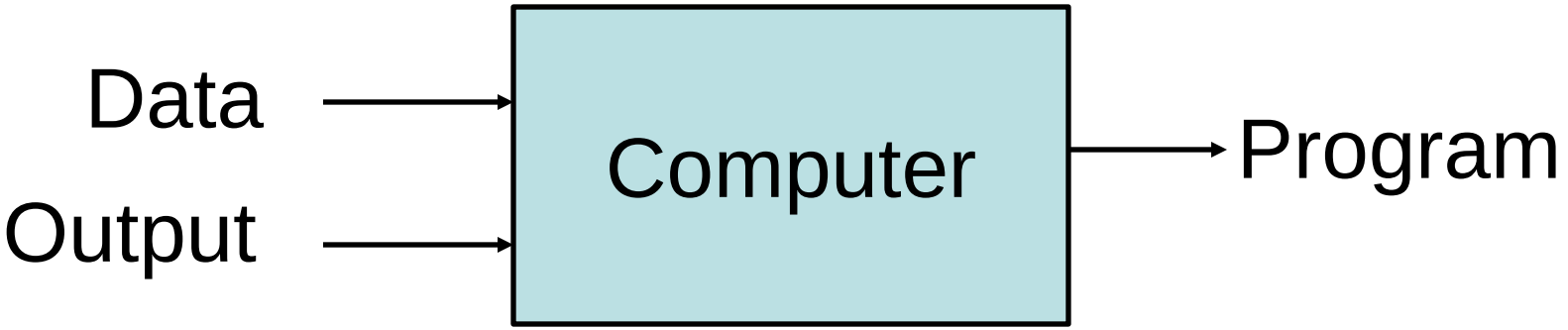
CONTINUE WITH LINEAR REGRESSION

TRAINING AND TESTING

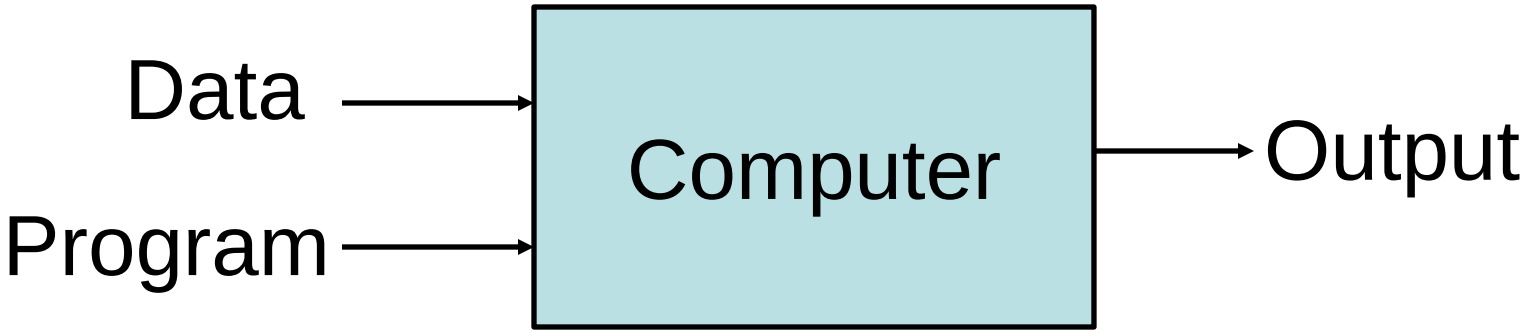
How does a prediction, supervised problem work?



- **Model gets built**



- **Use the model to predict new outputs**



There are many model options

Q: Which one do we choose?

There are many model options

Q: Which one do we choose?

Let’s choose the model that gives us the best performance

There are many model options

Q: Which one do we choose?

Let’s choose the model that gives us the best performance

Q: How do we measure performance? How well does it work?

There are many model options

Q: Which one do we choose?

Let’s choose the model that gives us the best performance

Q: How do we measure performance? How well does it work?

Can we use our dataset for an error estimate?

There are many model options

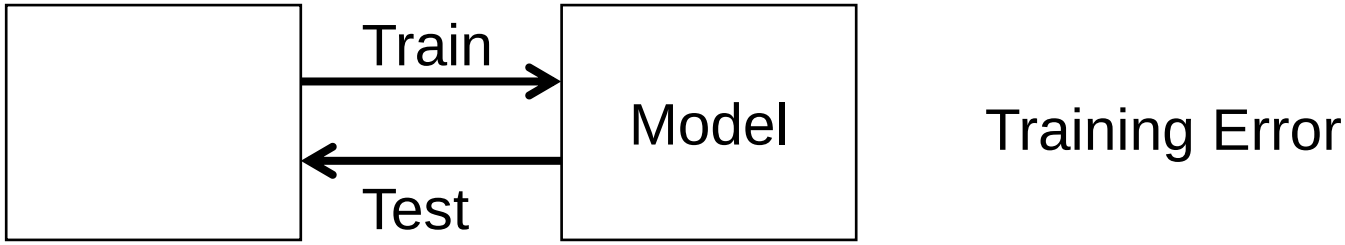
Q: Which one do we choose?

Let's choose the model that gives us the best performance

Q: How do we measure performance? How well does it work?

Can we use our dataset for an error estimate?

How would this work? Issues?



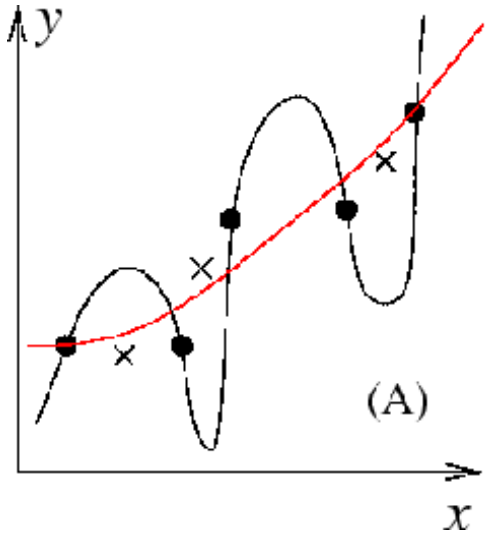
SUPERVISED LEARNING

Q: Are there any issues with training error?

Q: How small can we make our training error?

Q: Are there any issues with training error?
Q: How small can we make our training error?

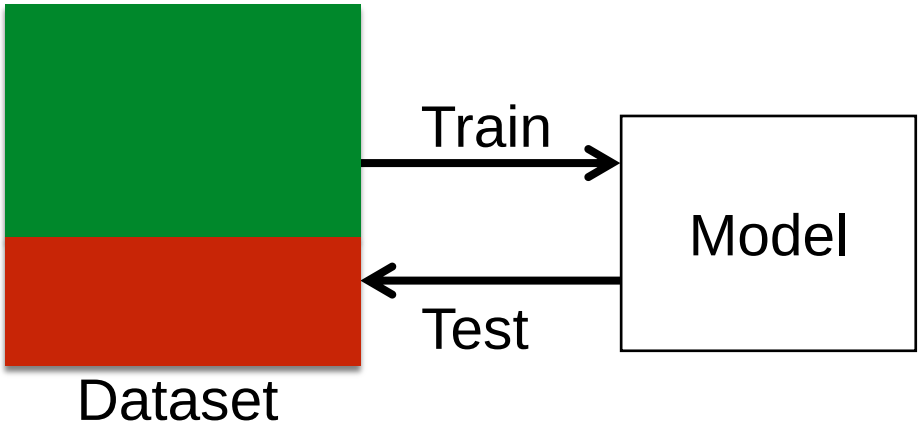
A: We can make the training error go to zero. We just need to memorize.



This is called over fitting

SUPERVISED LEARNING

Want performance on new observations. Data that we haven't seen



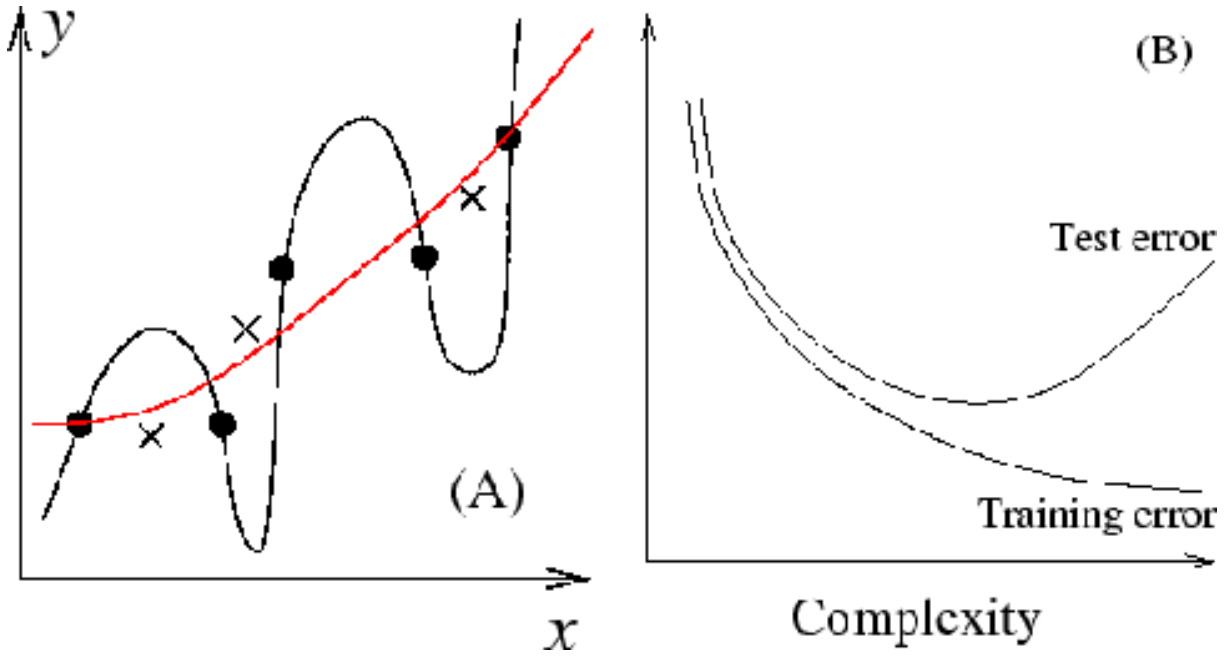
Training Error: Error on the training set

Test Error: Error on the test set

SUPERVISED LEARNING

In general:

- Training error goes to zero by adding complexity
- Test error goes down initially but then goes up (Overfitting)



K NEAREST NEIGHBOR

KNN: K-NEAREST NEIGHBOR

- A supervised learning algorithm
- Instance based learning
- Lazy learner
- Used for both regression and classification
- Voting and averaging

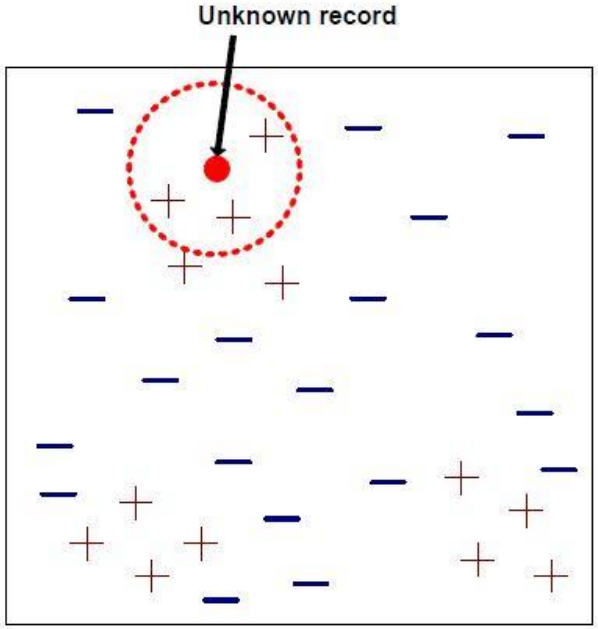
KNN: K-NEAREST NEIGHBOR

- A supervised learning algorithm
- Instance based learning
- Lazy learner
- Used for both regression and classification
- Voting and averaging

Q: How does this work?

KNN: K-NEAREST NEIGHBOR

- A supervised learning algorithm
- Instance based learning
- Lazy learner
- Used for both regression and classification
- Voting and averaging



Q: How does this work?

- Choose a k
- Calculate distances
- Average or vote

KNN: K-NEAREST NEIGHBOR

Advantages:

- Training is very fast
- Learn complex target functions
- Don’t lose information

Disadvantages:

- Slow at query time
- Lot’s of storage
- Easily fooled by irrelevant features and high dimensions