

An abstract graphic on the left side of the slide, featuring a sphere-like shape composed of a dense grid of intersecting red, green, and blue lines. The lines are curved and follow the contours of the sphere, creating a complex, woven pattern. The sphere is set against a dark gray background.

10-315

Introduction to ML

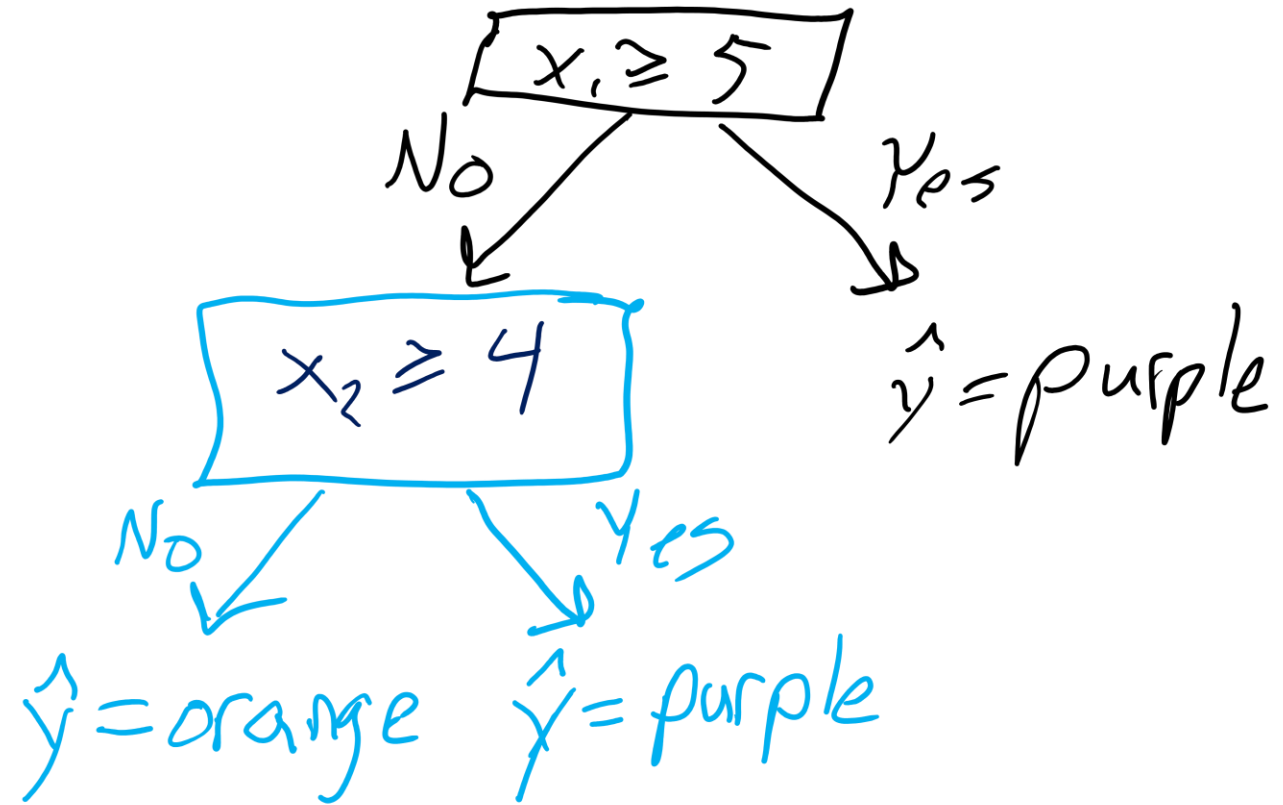
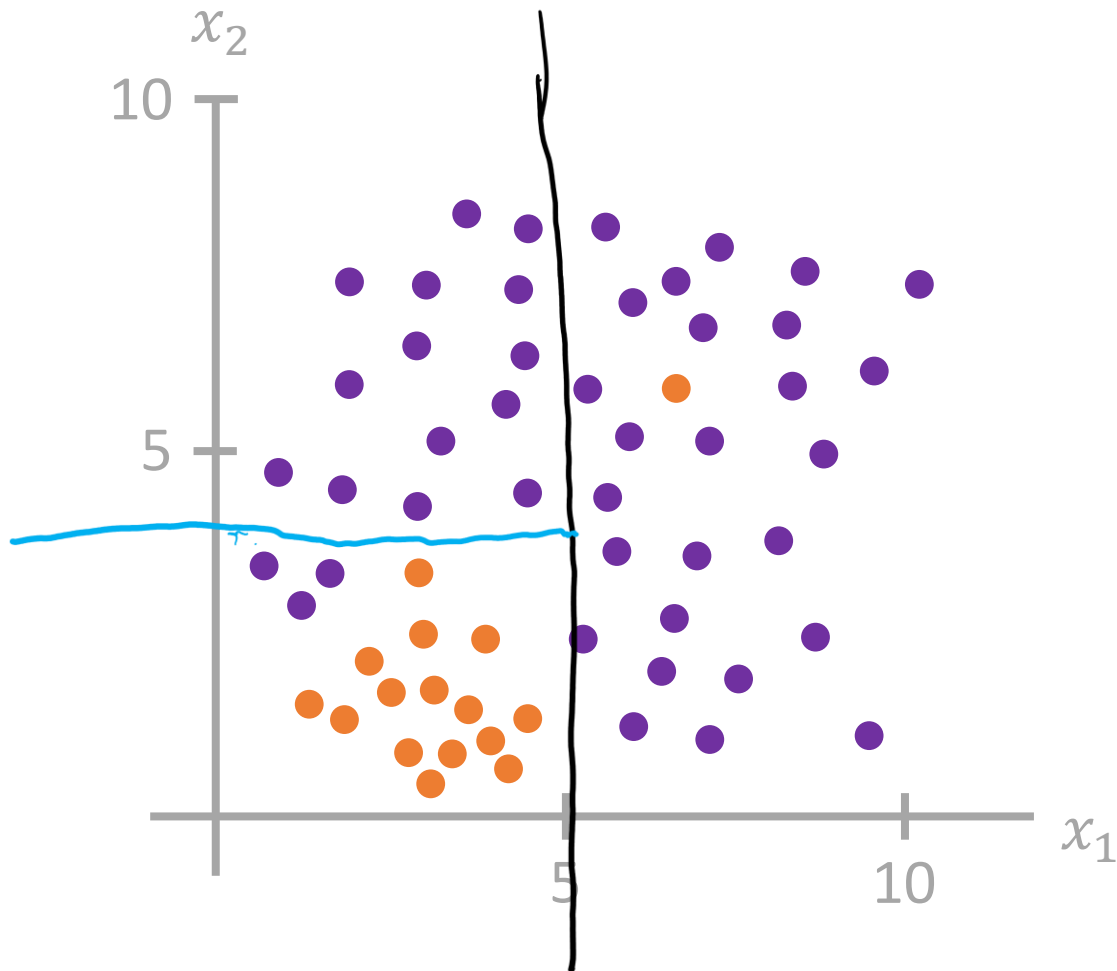
Model Selection

Instructor: Pat Virtue

Reminder: Decision Tree Worksheet

Consider input features $x \in \mathbb{R}^2$.

Draw a reasonable decision tree.



Poll 2

Decision tree generalization

Which of the following generalize best to unseen examples?

- A. Small tree with low training accuracy
- B. Large tree with low training accuracy
- C. Small tree with high training accuracy
- D. Large tree with high training accuracy

Underfitting and Overfitting

Underfitting occurs when model:

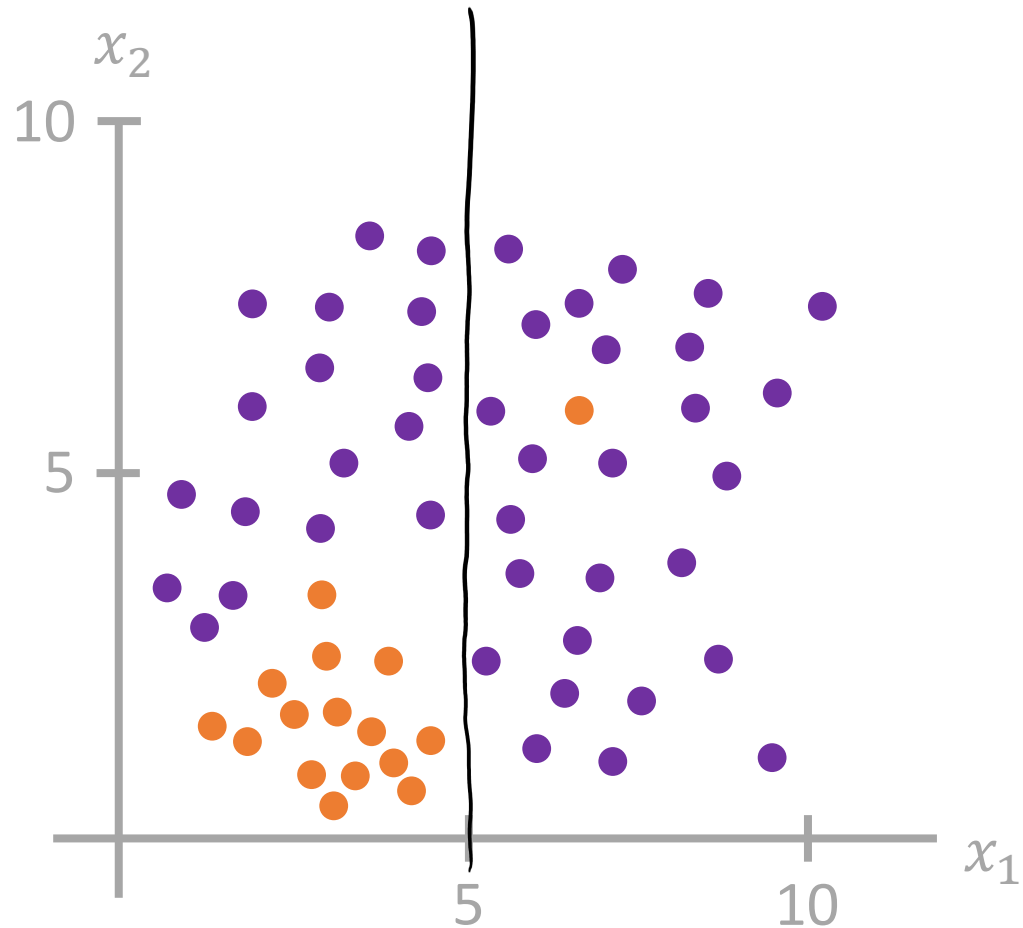
- is too simple
- can't capture the actual pattern of interest in the training dataset
- has too much inductive bias

Overfitting occurs when model:

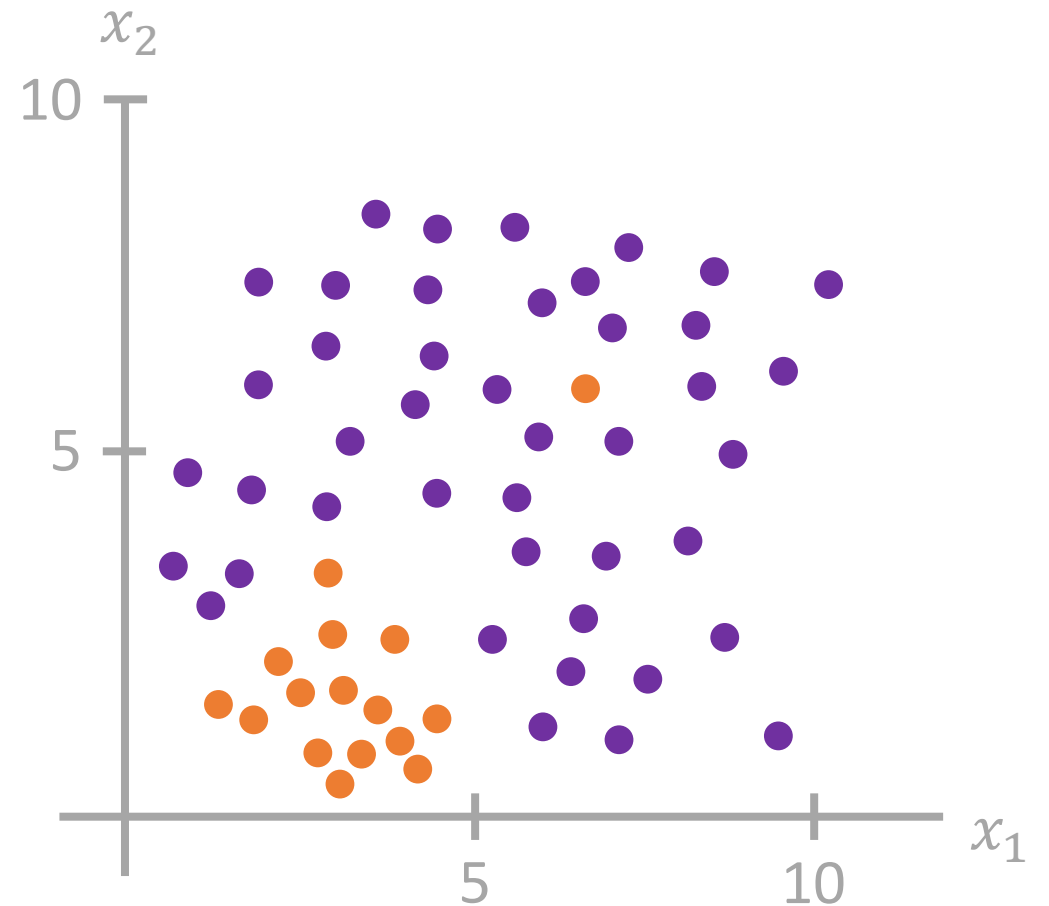
- is too complex
- fits noise or “outliers” in the training dataset as opposed to the actual pattern of interest
- doesn't have enough inductive bias pushing it to generalize

Underfitting and Overfitting: Classification

Underfitting

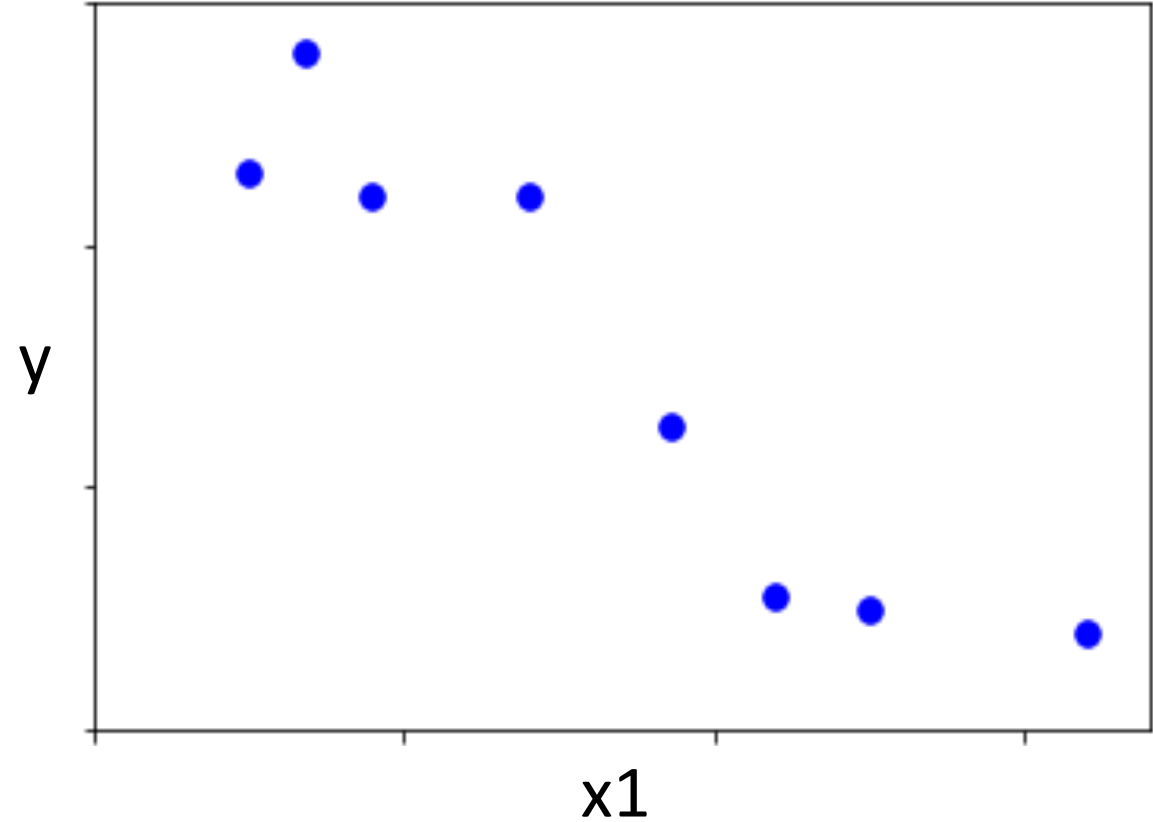


Overfitting

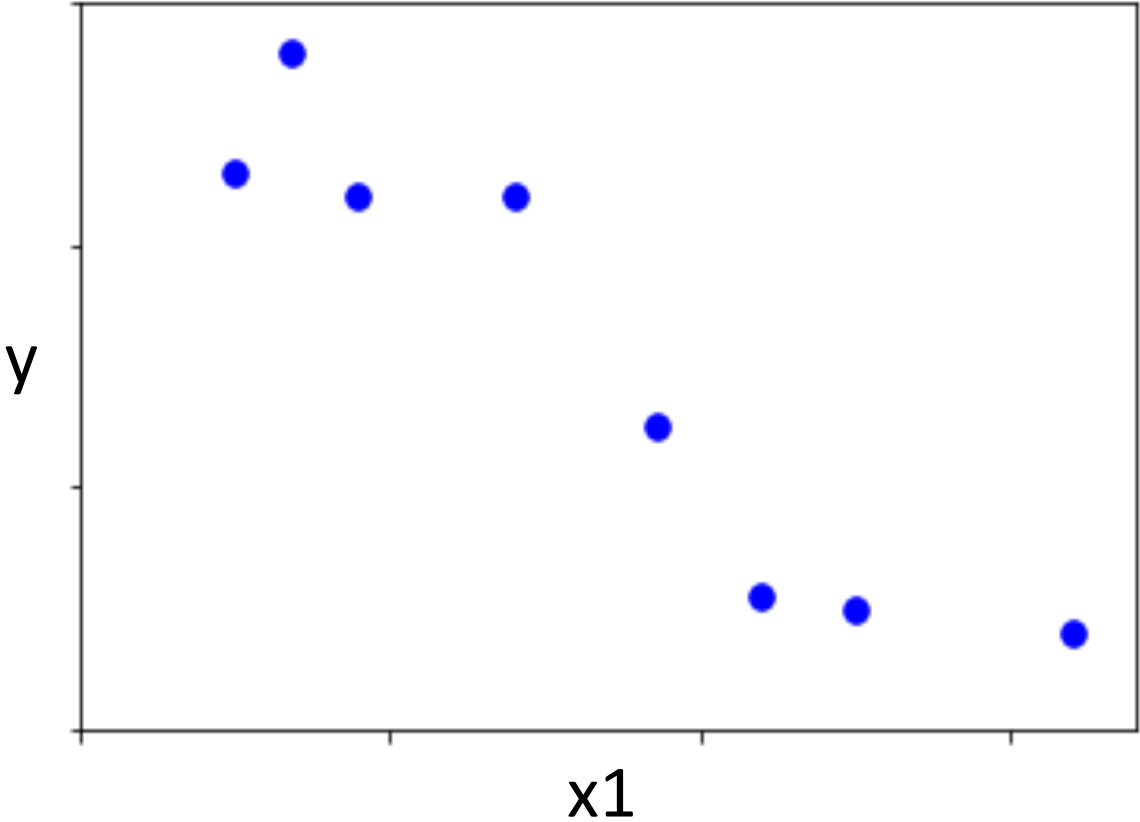


Underfitting and Overfitting: Regression

Underfitting

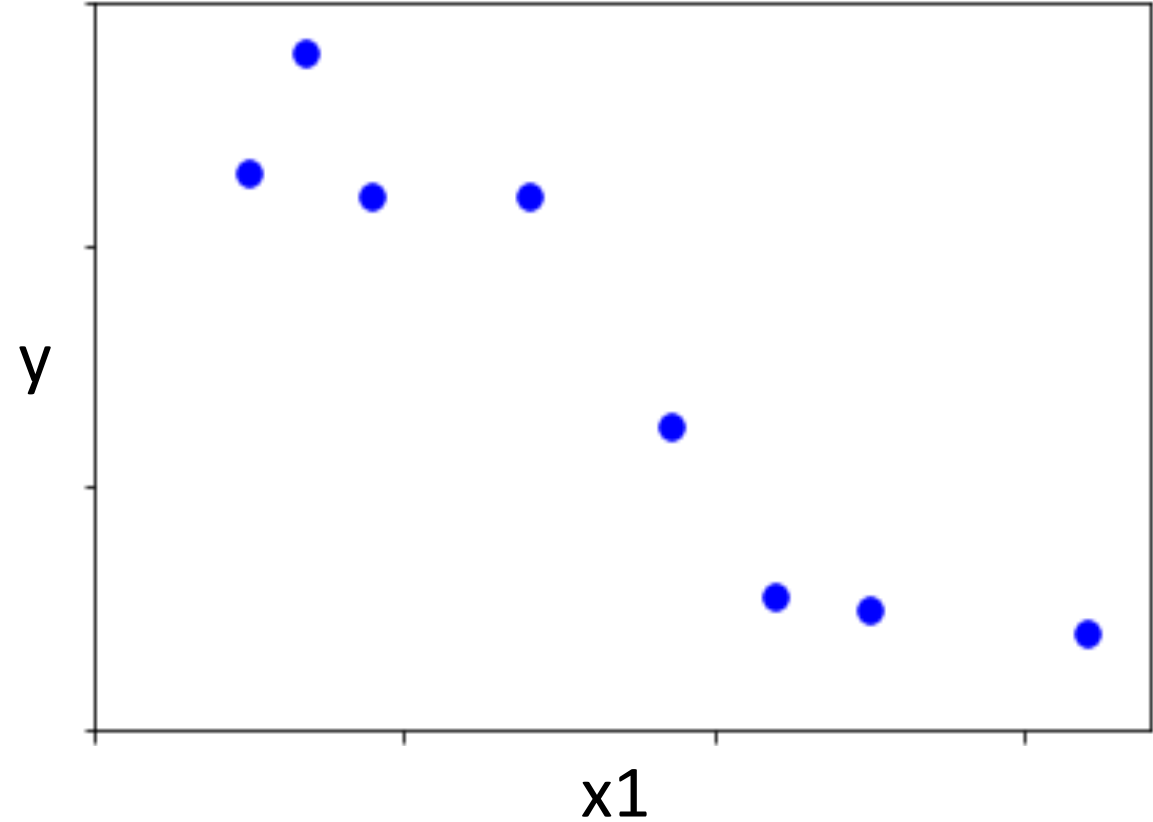


Overfitting

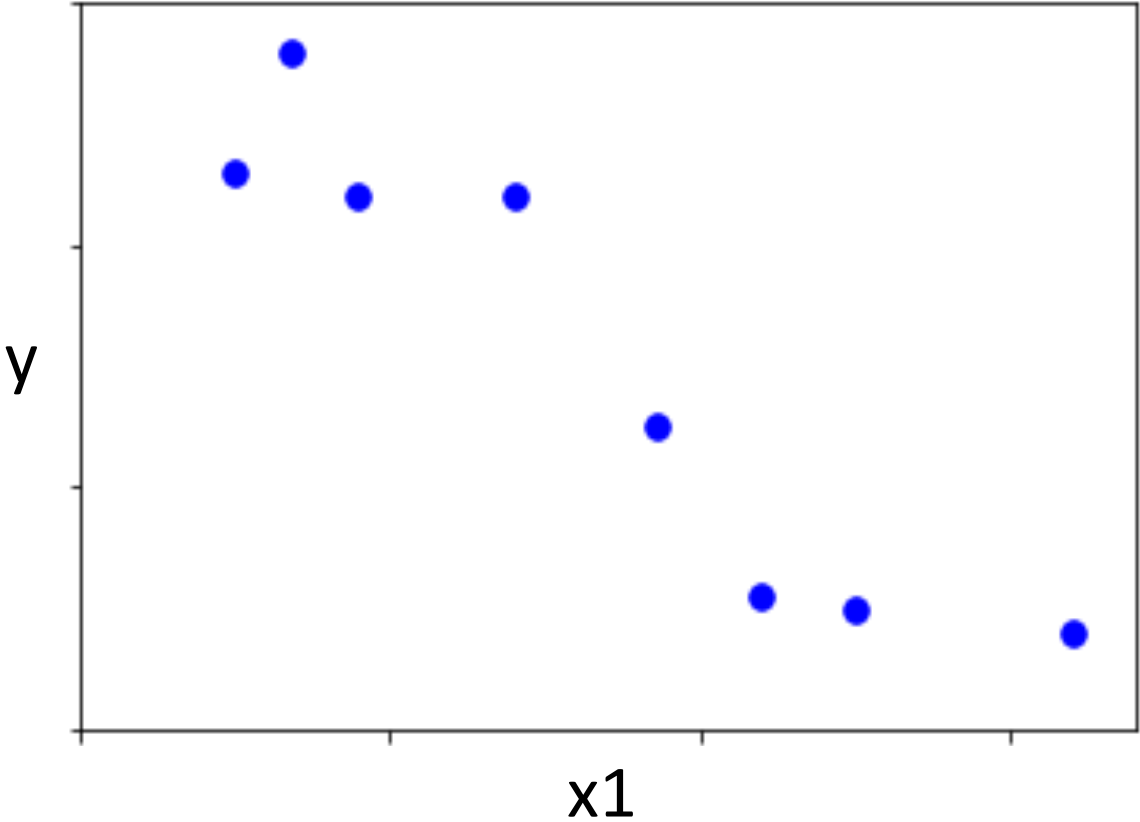


Underfitting and Overfitting: Regression

Underfitting



Overfitting



Reminder: Decision Trees

When do we stop (base case)?

When leaves are “pure”, i.e., output values are all the same

- Likely to overfit

Limit

- Tree depth
- Total number of leaves
- Splitting criteria threshold, e.g. splitting criteria $\leq \tau$
- Minimum number of datapoints in a leaf

But how do we choose all of those limits?!?

Answer: Model selection

Model selection is the process to choose the “best” among a set of (trained) models

Today

k-Nearest Neighbor

- 1-NN, k-NN (mostly notation)
- Practical details

Underfitting and Overfitting

- Decision Tree stopping criteria
- Classification and regression examples

Model Selection

- Define terms
- Stress importance
- Cross-validation techniques



Model Selection

Model Selection

WARNING:

- In some sense, our discussion of model selection is premature.
- The models we have considered thus far are fairly simple.
- The models and the many decisions available to the data scientist wielding them will grow to be much more complex than what we've seen so far.

Model Selection

Statistics

- *Def:* a **model** defines the data generation process (i.e. a set or family of parametric probability distributions)
- *Def:* **model parameters** are the values that give rise to a particular probability distribution in the model family
- *Def:* **learning** (aka. estimation) is the process of finding the parameters that best fit the data
- *Def:* **hyperparameters** are the parameters of a prior distribution over parameters

Machine Learning

- *Def:* (loosely) a **model** defines the hypothesis space over which learning performs its search
- *Def:* **model parameters** are the numeric values or structure selected by the learning algorithm that give rise to a hypothesis
- *Def:* the **learning algorithm** defines the data-driven search over the hypothesis space (i.e. search for good parameters)
- *Def:* **hyperparameters** are the tunable aspects of the model, that the learning algorithm does not select

Model Selection

Example: Decision Tree

- model = set of all possible trees, possibly restricted by some hyperparameters (e.g. max depth)
- parameters = structure of a specific decision tree
- learning algorithm = ID3, CART, etc.
- hyperparameters = max-depth, threshold for splitting criterion, etc.

Machine Learning

- *Def:* (loosely) a **model** defines the hypothesis space over which learning performs its search
- *Def:* **model parameters** are the numeric values or structure selected by the learning algorithm that give rise to a hypothesis
- *Def:* the **learning algorithm** defines the data-driven search over the hypothesis space (i.e. search for good parameters)
- *Def:* **hyperparameters** are the tunable aspects of the model, that the learning algorithm does not select

Model Selection

Example: k-Nearest Neighbors

- model = set of all possible nearest neighbors classifiers
- parameters = none
(KNN is an instance-based or non-parametric method)
- learning algorithm = for naïve setting, just storing the data
- hyperparameters = k , the number of neighbors to consider

Machine Learning

- *Def:* (loosely) a **model** defines the hypothesis space over which learning performs its search
- *Def:* **model parameters** are the numeric values or structure selected by the learning algorithm that give rise to a hypothesis
- *Def:* the **learning algorithm** defines the data-driven search over the hypothesis space (i.e. search for good parameters)
- *Def:* **hyperparameters** are the tunable aspects of the model, that the learning algorithm does *not* select

Model Selection

Statistics

- Def: a **model** defines the data generation process (i.e. a set or family of probability distributions)
- Def: **model parameters** are the parameters that give rise to a particular probability distribution in the model family
- Def: **learning** (aka. estimation) is the process of finding the parameters that best fit the data
- Def: **hyperparameters** are the parameters of a prior distribution over parameters

Machine Learning

- Def: (loosely) a **model** defines the hypothesis space in which learning performs its search
- Def: **model parameters** are the numeric values of the model structure selected by the learning algorithm that give rise to a hypothesis
- Def: the **learning algorithm** defines the data-driven search over the hypothesis space (i.e. search for good parameters)
- Def: **hyperparameters** are the tunable aspects of the model, that the learning algorithm does not select

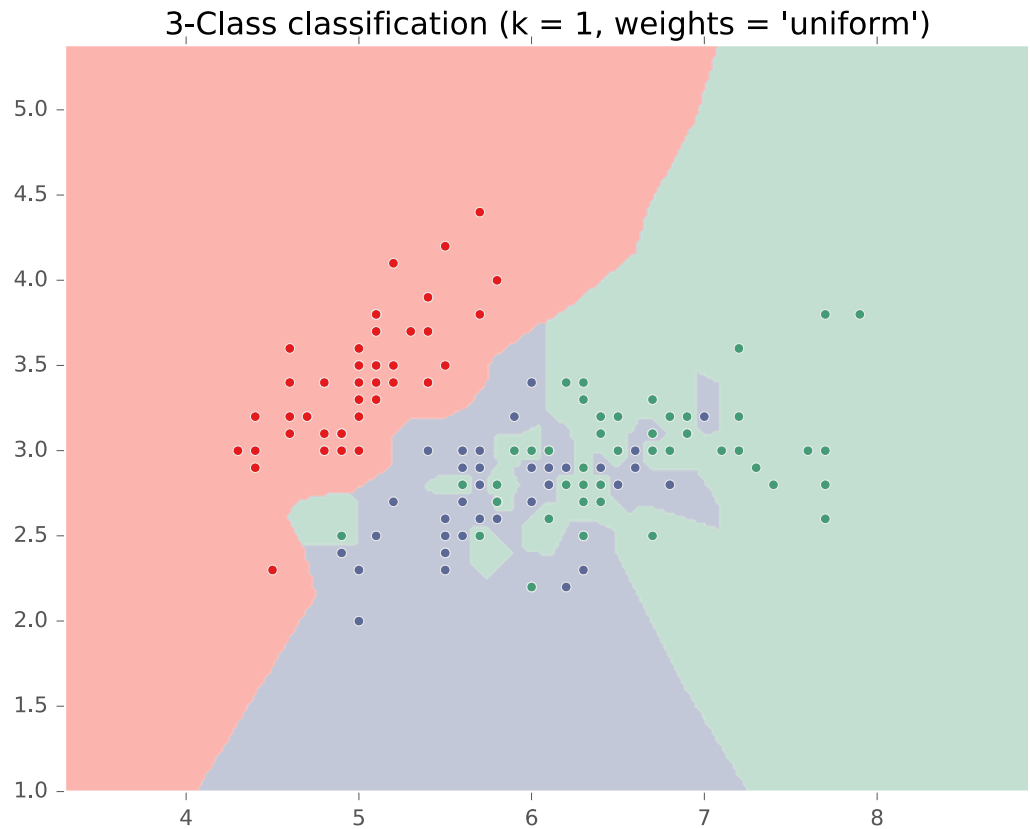
If “learning” is all about picking the best **parameters** how do we pick the best **hyperparameters**?

Model Selection

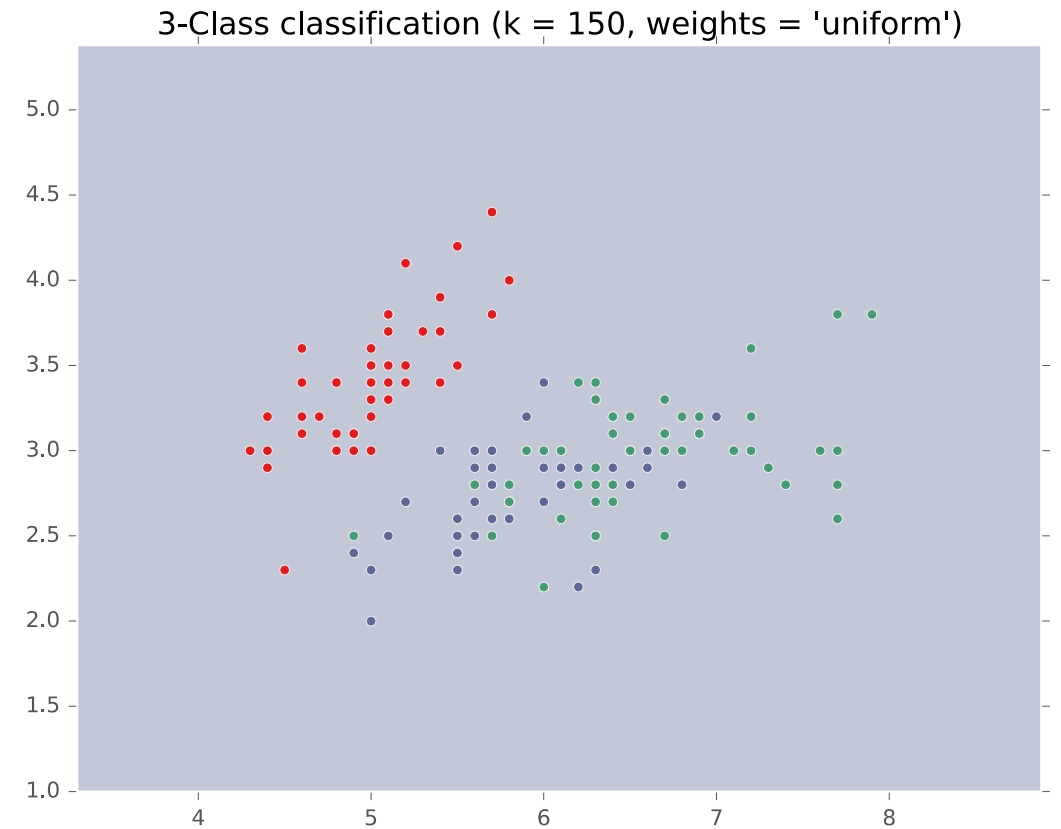
- Two very similar definitions:
 - Def: **model selection** is the process by which we choose the “best” model from among a set of candidates
 - Def: **hyperparameter optimization** is the process by which we choose the “best” hyperparameters from among a set of candidates (**could be called a special case of model selection**)
- **Both** assume access to a function capable of measuring the quality of a model
- **Both** are typically done “outside” the main training algorithm --- typically training is treated as a black box

Special Cases of k-NN

k=1: Nearest Neighbor

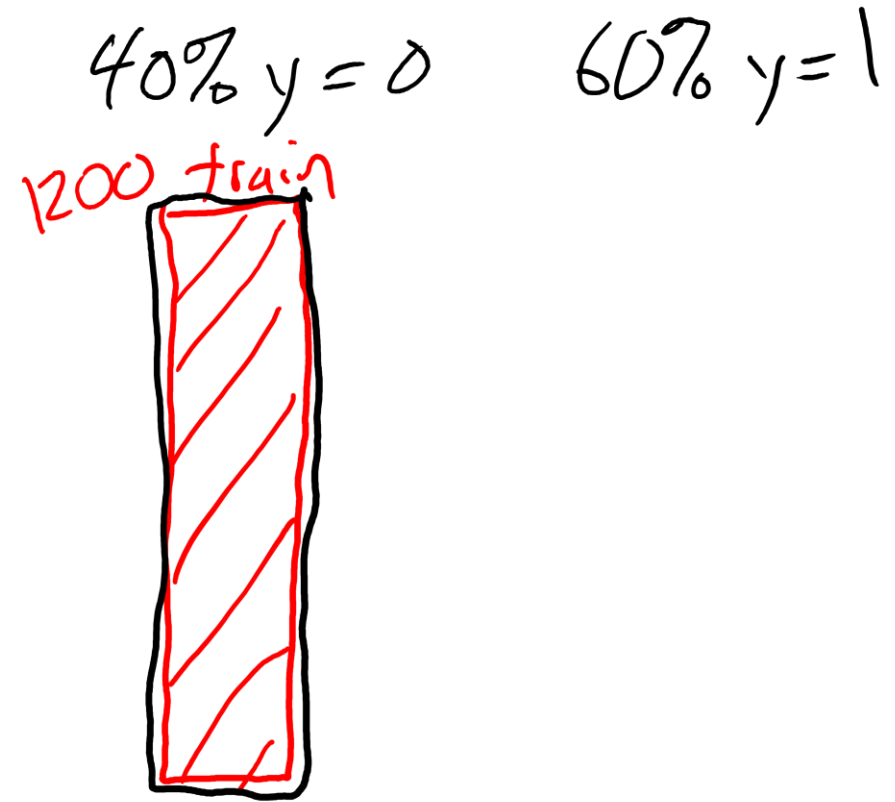
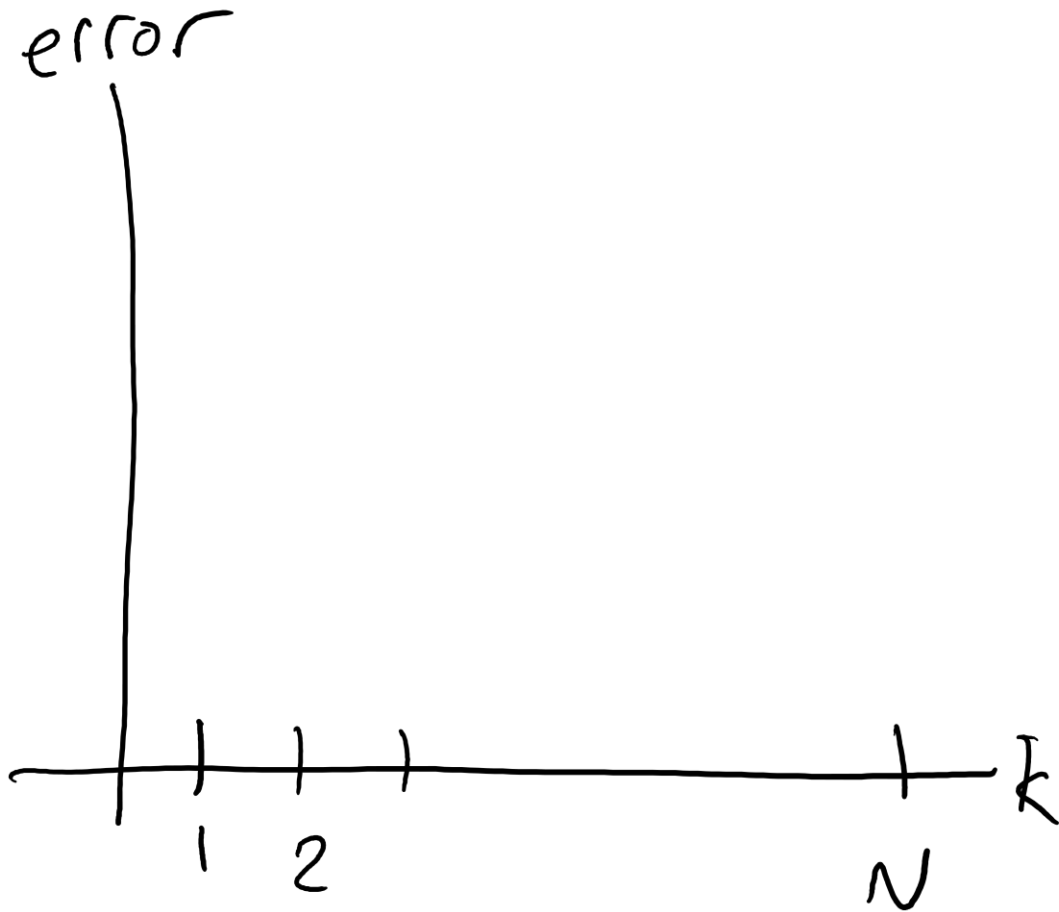


k=N: Majority Vote

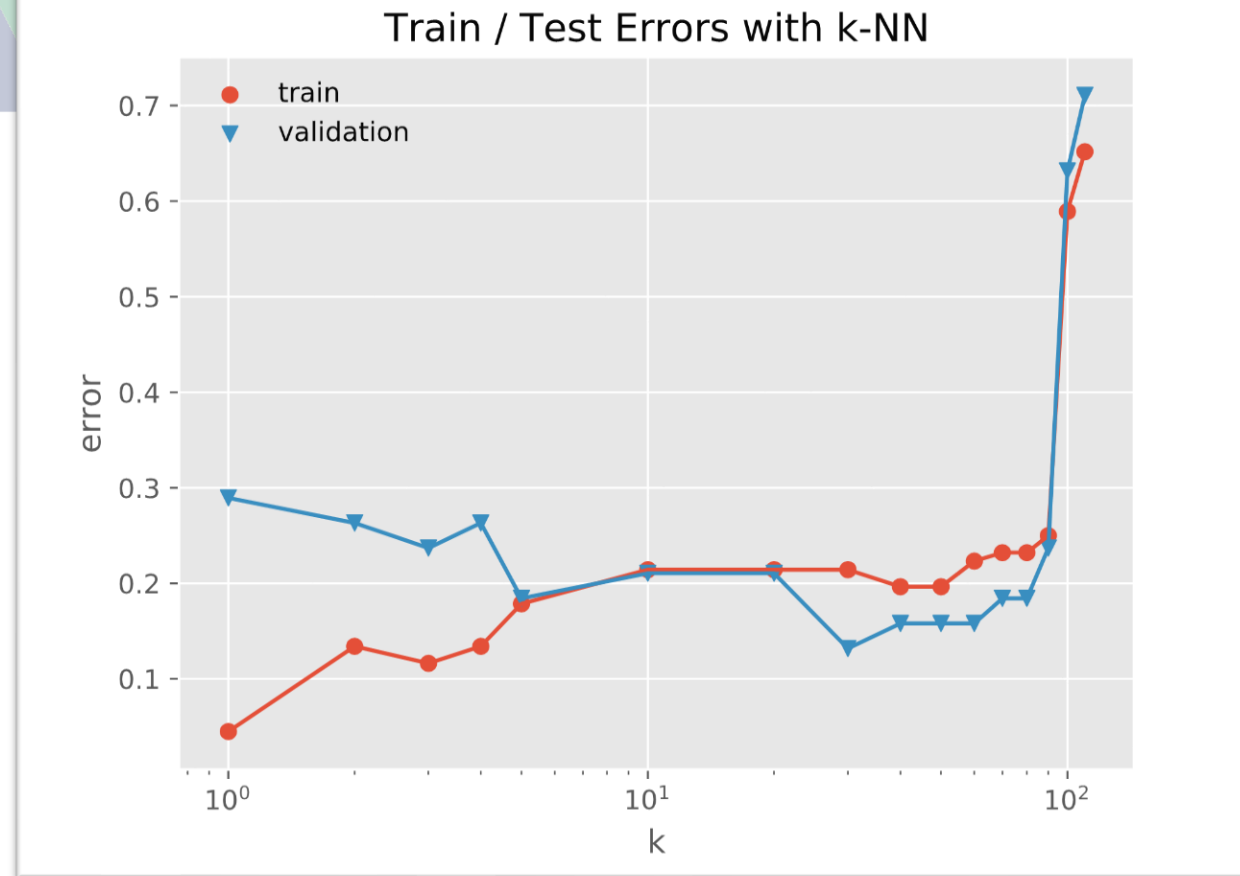
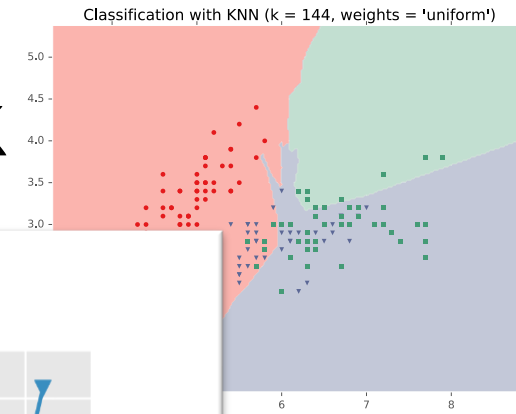
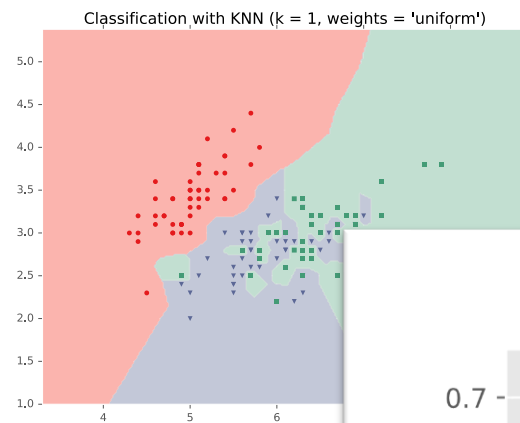


Example of Hyperparameter Optimization

Choosing k for k -NN

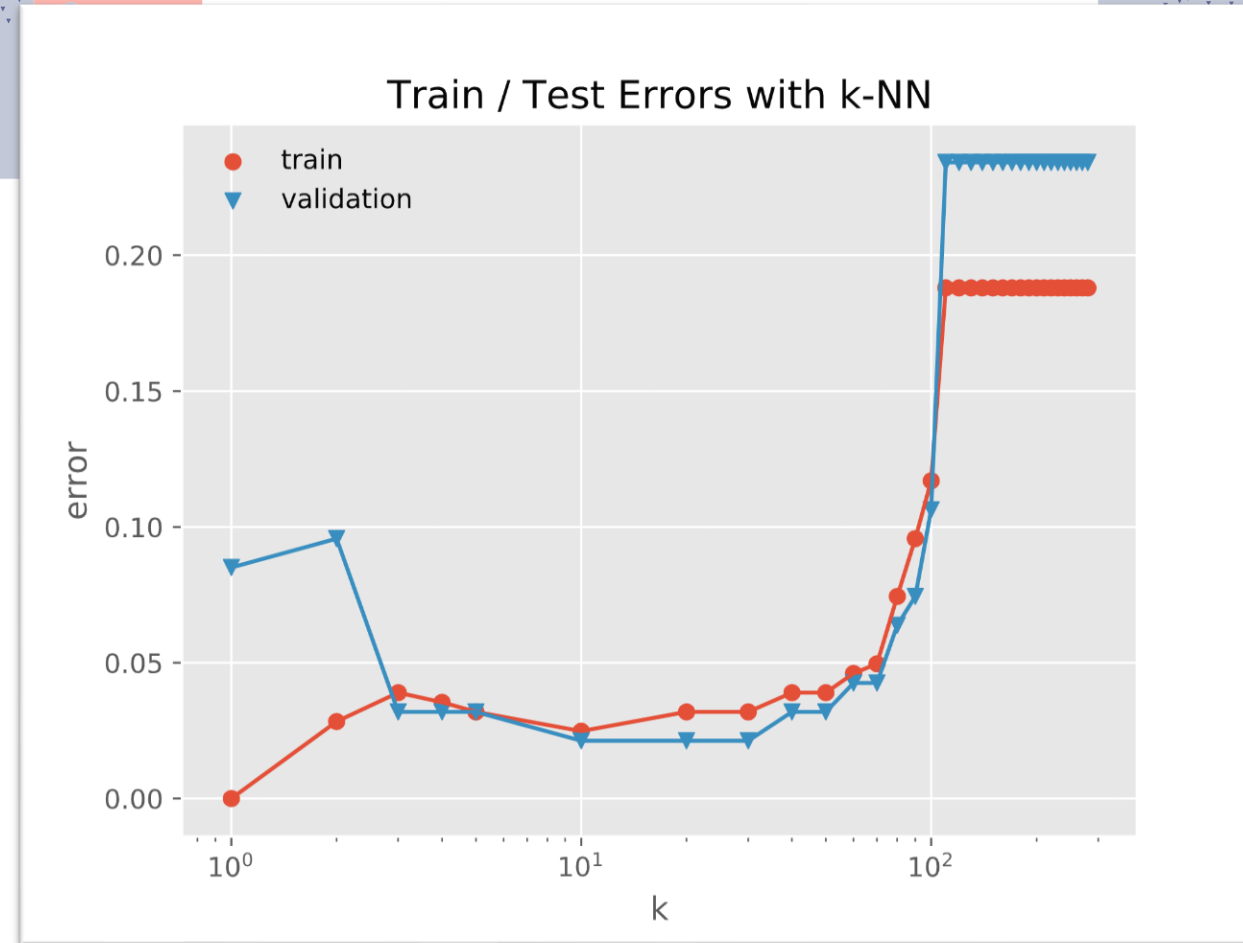
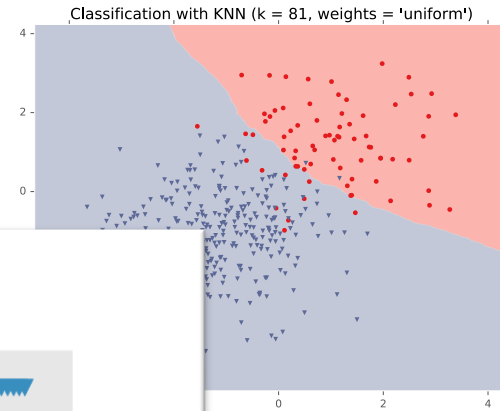
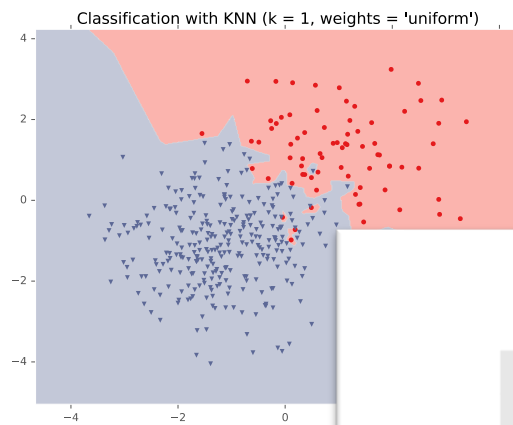


k-NN: Choosing k



Fisher Iris Data: varying the value of k

k-NN: Choosing k



Gaussian Data: varying the value of k

Cross-validation

Why do we need cross-validation?

- Choose hyperparameters
- Choose technique
- Help make any choices beyond our parameters

But now, we have another choice to make!

- How do we split training and validation?

Trade-offs

- More held-out data, better meaning behind validation numbers
- More held-out data, less data to train on!

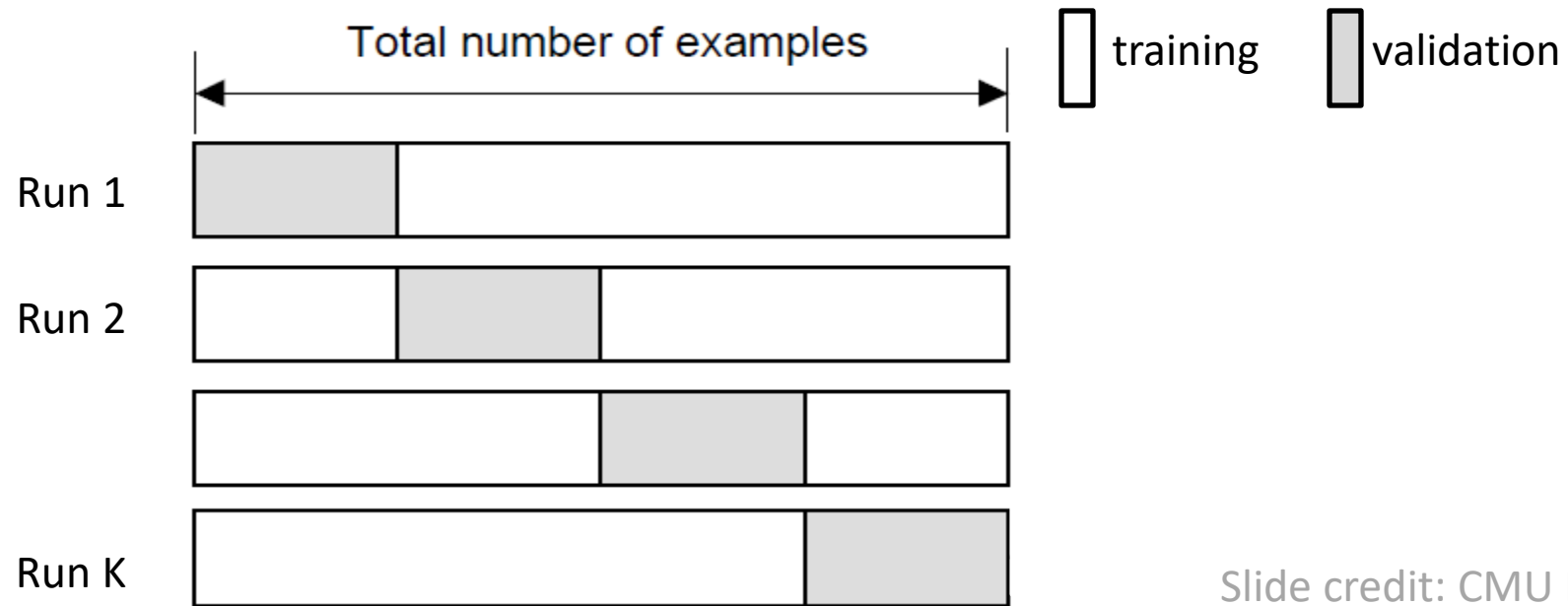
Cross-validation

K-fold cross-validation

Create K-fold partition of the dataset.

Do K runs: train using K-1 partitions and calculate validation error on remaining partition (rotating validation partition on each run).

Report average validation error

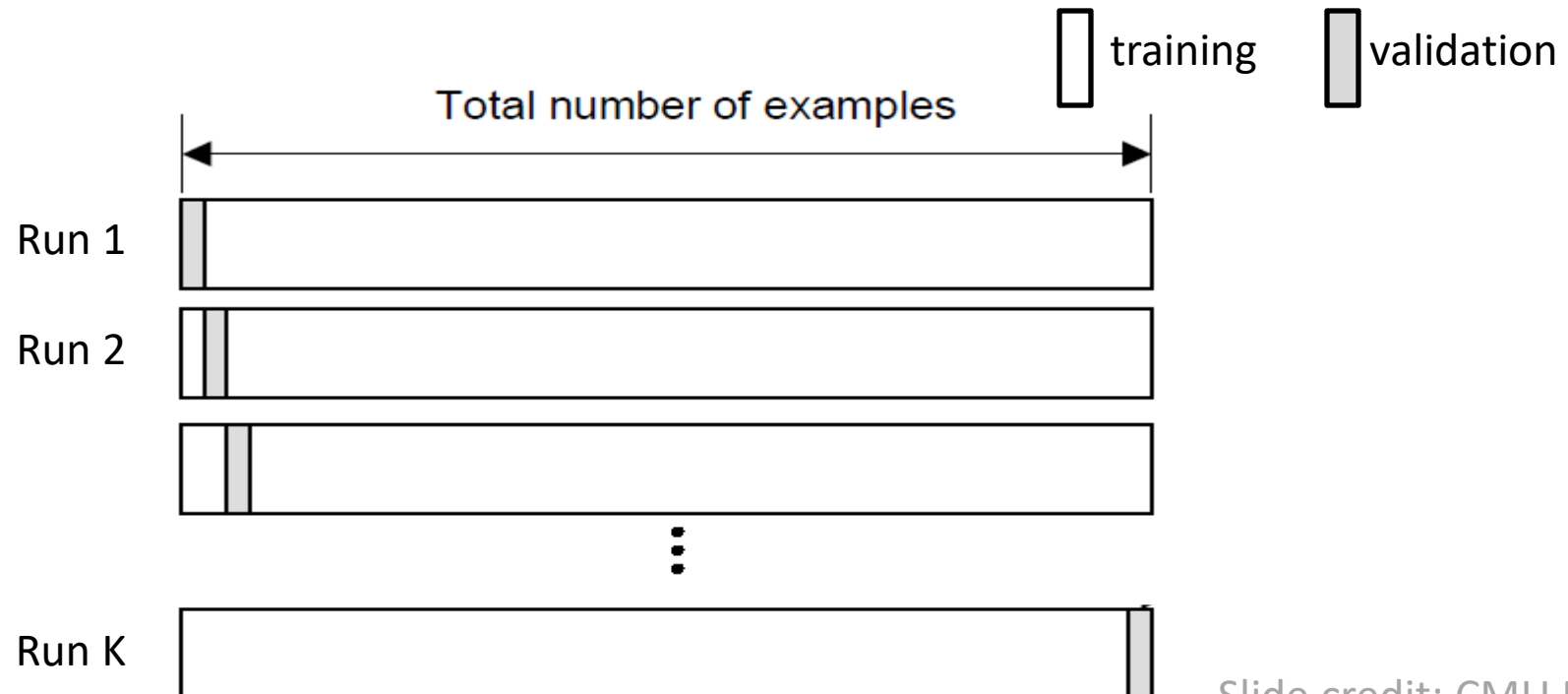


Cross-validation

Leave-one-out (LOO) cross-validation

Special case of K-fold with $K=N$ partitions

Equivalently, train on $N-1$ samples and validate on only one sample per run for N runs



Cross-validation

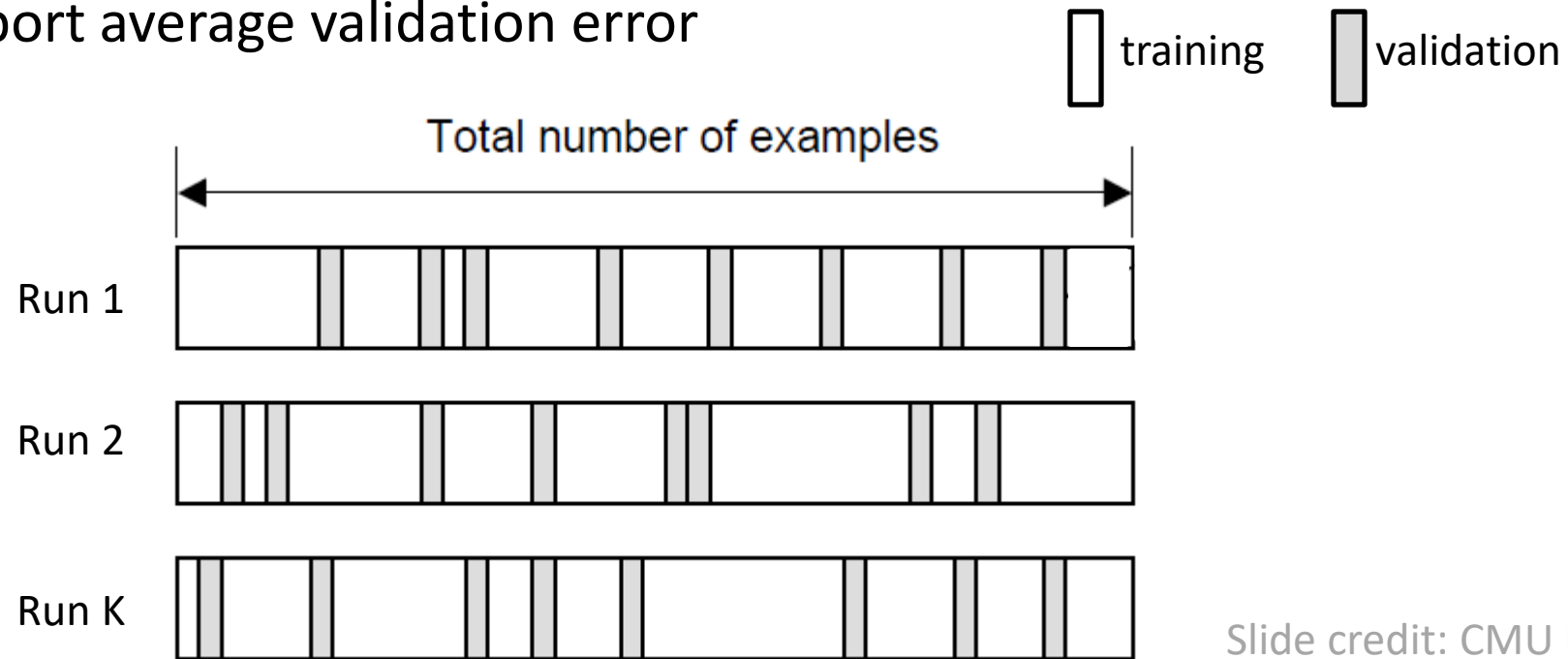
Random subsampling

Randomly subsample a fixed fraction αN ($0 < \alpha < 1$) of the dataset for validation.

Compute validation error with remaining data as training data.

Repeat K times

Report average validation error



Poll 3

Say you are choosing amongst 7 discrete values of a decision tree *mutual information threshold*, and you want to do K=5-fold cross-validation.

How many times do I have to train my model?

- A. 1
- B. 5
- C. 7
- D. 12
- E. 35
- F. 5^7

Experimental Design

	Input	Output	Notes
Training	<ul style="list-style-type: none">• training dataset• hyperparameters	<ul style="list-style-type: none">• best model parameters	We pick the best model parameters by learning on the training dataset for a fixed set of hyperparameters
Hyperparameter Optimization	<ul style="list-style-type: none">• training dataset• validation dataset	<ul style="list-style-type: none">• best hyperparameters	We pick the best hyperparameters by learning on the training data and evaluating error on the validation error
Testing	<ul style="list-style-type: none">• test dataset• hypothesis (i.e. fixed model parameters)	<ul style="list-style-type: none">• test error	We evaluate a hypothesis corresponding to a decision rule with fixed model parameters on a test dataset to obtain test error