Microsoft: DAT210x Programming with Python for Data Science

Bookmarks

- Start Here
- 1. The Big Picture
- 2. Data And Features
- 3. Exploring Data
- 4. Transforming Data
- 5. Data Modeling
- **6. Data Modeling II**

Lecture: SVC
Quiz

Lab: SVC
Lab

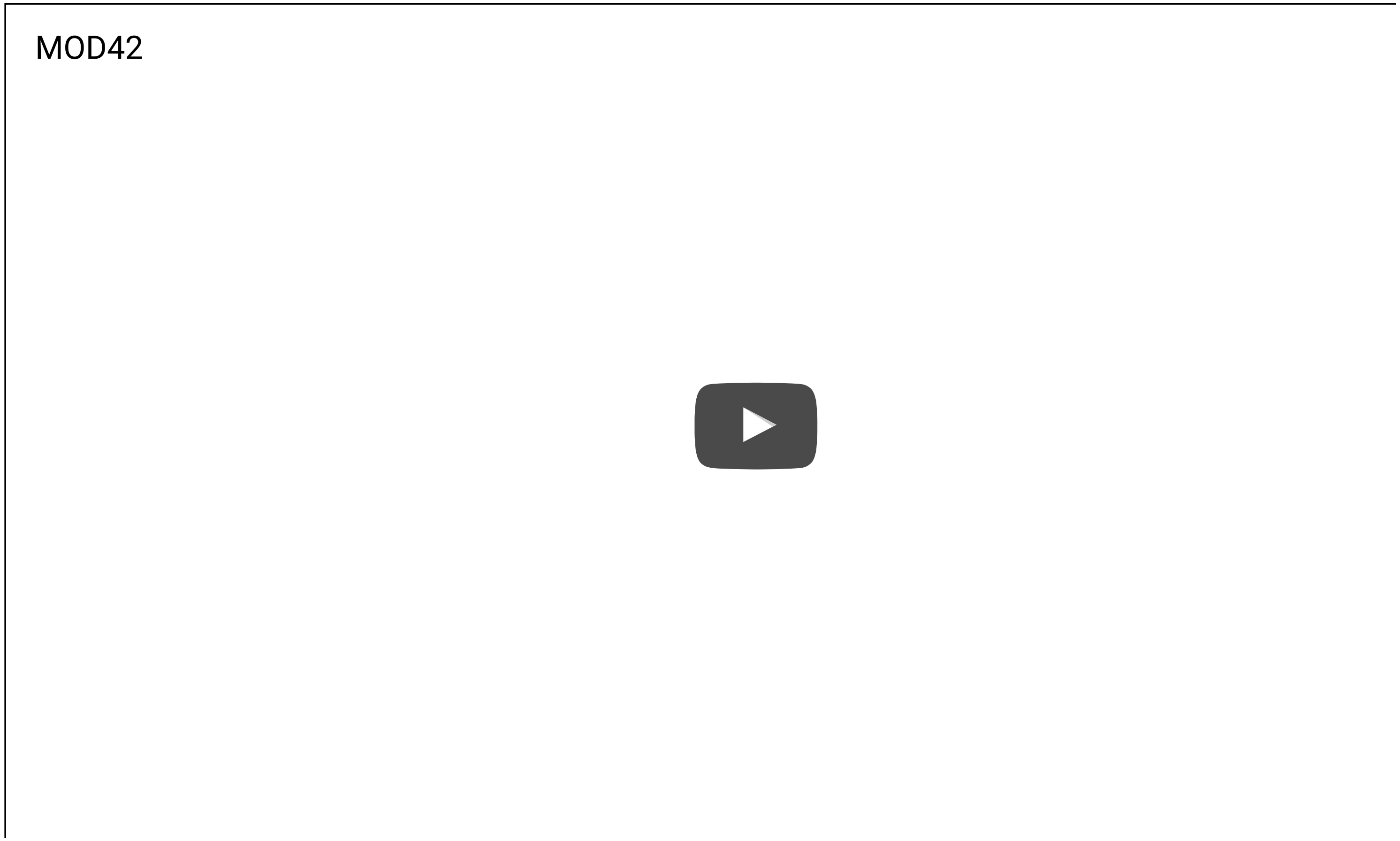
Lecture: Decision Trees
Quiz

Lab: Decision Trees
Lab

6. Data Modeling II > Lecture: Random Forest > Video

Bookmark

Random Forest



Lecture: Random Forest
Quiz

Dive Deeper

0:00 / 1:36 1.0x

Download video Download transcript .srt

Decision trees are a wonderful tool. They're easy to understand conceptually, fast to train and test, and offer a good level of accuracy. In Trevor Hastie's *The Elements of Statistical Learning (2008, 2nd ed.)*, he mentions decision trees "come closest to meeting the requirements for serving as an off-the-shelf procedure for data mining." Some of the advantages decision trees offer include:

- Unlike SVMs, the accuracy of a DTree doesn't decrease when you include irrelevant features
- Unlike KNeighbors, both training and predicting with a DTree are relatively fast operations
- Unlike PCA / IsoMap, DTrees are invariant to monotonic feature scaling and transformations
- Moreover, a trained DTree model is readily human inspectable

As Trevor Hastie indicated, under many circumstances, they're really a miracle machine learning algorithm, but with an Achilles heel—if not *carefully* pruned, decision trees will deep-learn irregular patterns and data outliers. They do this so well that they rapidly overfit the training set, resulting in excellent training data recall... but poor predictive abilities. This is exactly the problem random forests aim to solve.

When we send the first person to Mars in the future, it's going to be one way trip... so that person will need to have as much knowledge as possible. How to grow food, how to read martian weather patterns, various facets of engineering, solar system navigation, modes of keeping entertained, etc. To help that person prepare, the agency overseeing the journey will hire the *top* people from a broad range of fields to conduct intensive coaching. But is one person truly capable of learning *everything* they need to know, at the depth of complexity needed? What if that person has a particular bias towards a field? Their depth of knowledge in that area will likely be more than in fields they didn't care for, creating imbalance.

Being aware of this, the organizing agency will likely hedge their bets by sending a **team** of explorers to Mars, not just one. Even though it's more expensive for them due to the added weight and food requirements, they know it's worth it, since each person will have generic knowledge about all topics; but specific knowledge about their own area of interest.

This is exactly how random forest work. A single decision tree, tasked to learn a dataset might not be able to perform well due to the outliers, and the breadth and depth complexity of the data. So instead of relying on a single tree, random forests rely on a forest of cleverly grown decision trees. Each tree within the forest is allowed to become highly specialized in a specific area, but still retains some general knowledge about most areas. When a random forest classifier, it is actually each tree in the forest working together to cast votes on what label they think a specific sample should be assigned.