Decision Trees: Conclusion

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

Decision Trees

- **Decision trees are the single most popular data mining tool:**
  - Easy to understand
  - Easy to implement
  - Easy to use
  - Computationally cheap
  - It's possible to get in trouble with overfitting
  - **They do classification as well as regression!**

Learning Ensembles

- **Learn multiple trees and combine their predictions**
  - Gives better performance in practice
- **Bagging:**
  - Learns multiple trees over independent samples of the training data
  - Predictions from each tree are averaged to compute the final model prediction

Bagged Decision Trees

- **Model construction for bagging in PLANET**
  - When tree induction begins at the root, nodes of all trees in the bagged model are pushed onto the **MRQ** queue
  - Controller does tree induction over dataset samples
    - Queues will contain nodes belonging to many different trees instead of a single tree
- **How to create random samples of D*?**
  - Compute a hash of a training record's id and tree id
  - Use records that hash into a particular range to learn a tree
  - This way the same sample is used for all nodes in a tree
  - **Note:** This is sampling **D*** without replacement (but samples of **D*** should be created with replacement)

SVM vs. DT

- **SVM**
  - **Classification**
    - Usually only 2 classes
  - **Real valued features** (no categorical ones)
  - **Tens/hundreds of thousands of features**
  - Very **sparse features**
  - **Simple decision boundary**
    - No issues with overfitting
- Example applications
  - Text classification
  - Spam detection
  - Computer vision

- **Decision trees**
  - **Classification & Regression**
    - Multiple (~10) classes
  - **Real valued and categorical features**
  - **Few (hundreds) of features**
  - Usually **dense features**
  - **Complicated decision boundaries**
    - Overfitting! Early stopping
- Example applications
  - User profile classification
  - Landing page bounce prediction