Lesson Plan 4/26

Thursday, April 26, 2018 7:48 AM

Admin
- HW3 Due date moved
- HW3 Part 2 released

HDSC
- Ursa labs, shared infrastructure for data science: https://ursalabs.org/

Random Forests

Classifier Evaluation

Cross-validation

Decision Trees

— Recursive Partioning

— Stopping rule:
stop if only 1 training example is left

→ Very deep trees

→ Overfit training set

→ Prone trees

⇒ Depth of the tree affects predictive performance

⇒ Trees are highly unstable

Random forests

Ensemble Method:

→ Build Multiple classifiers

→ Aggregate predictions

Classifier: Decision Tree (deep, no pruning)

Ensure diversity in predictors:

→ Bootstrapping training set

→ 'Random' split selection

Bootstrapping:

→ # of observations is the same as the original dataset

→ Distribution of attributes $P(Y \mid X)$ is very similar to original dataset

Tree bording:

→ Randomly select subset of

predictors to consider as splits

Aggregation:

Regression: average of regression tree predictors

Classification: majority class

→ Is RF better than a single tree?

→ Does the # trees in the forest affect performance?

Classifier Performance:

Confusion Matrix

	Observed 0 (negative)	Observed 1 (positive)
predicted 0	TN	FN
1	FP	TP

"Imbalanced data set"

False Positive

False Negative

Trade-off

precision vs recall

TPR vs. FPR

Logistic Regression

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} > 0 \quad \text{then predict } 1$$

Change (decision rule:)

$$P(Y=1|X) \quad \cdot \quad \cdot \quad \cdot$$

$\log \frac{P(Y=1|x)}{P(Y=0|x)} > 1$ then predict 1

↑↑ precision

↓↓ recall

Receiver Operator Curve