Business Analytics

Classification

Instructor: Hrant Davtyan
Course: Business Analytics, Fall, MSSM, 2018

Content

1. Overview: model types
2. Intro to Classification

3. Model Fitting
4. Model Evaluation
5. Model Diagnostics
6. Model Selection

7. Logistic Regression
8. Decision Tree Classification

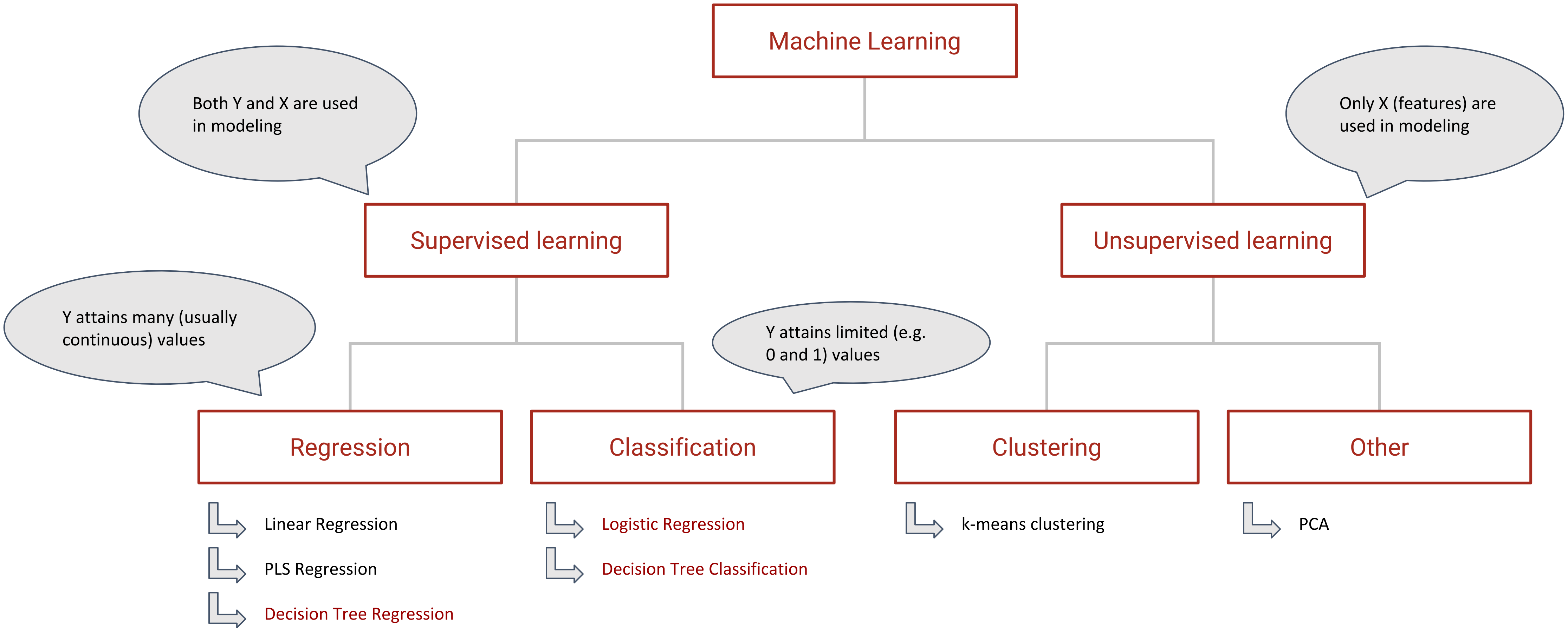
Content

1. Overview: model types
2. Intro to Classification

3. ~~Model fitting~~
4. Model Evaluation
5. ~~Model Diagnostics~~
6. Model Selection

7. Logistic Regression
8. Decision Tree Classification

Overview


Machine Learning
Both Y and X are used in modeling
Supervised learning
Only X (features) are used in modeling
Unsupervised learning
Y attains many (usually continuous) values
Y attains limited (e.g. 0 and 1) values
Regression
Classification
Clustering
Other
Linear Regression
PLS Regression
Decision Tree Regression
Logistic Regression
Decision Tree Classification
k-means clustering
PCA

Intro to Classification

Introduction to Classification

1. Classification models provide the opportunity to predict the class of target (dependent) variable Y using features (independent variables) in the dataset.
2. When Y attains only 2 values (simple case) we have a binary classification problem. Examples:
   - Predicting whether customer will buy (1) or not (0) a certain product
   - Understanding the factors affecting employee's decision to stay (0) or leave (1)
3. When Y attains more than 2 values (general case) we have a multiclass classification problem. Examples:
   - Predicting what type of car the customer will buy: Hatchback (0), Minivan (1) or other (2)
   - Predicting the winner party in elections

Introduction to Classification (cont'd)

1. Regression models assume that the target variable Y is continuous, thus they are not useful for classification. For example, if we want to predict employee turnover using regression model, it is mathematically possible to estimate Y=1.5 value when the highest value Y can attain is 1 (employee leaves).
2. Thus, we need new type of models that will only predict values that are in the [0,1] interval.
3. It is preferred not to predict turnover alone, but also assign probabilities to prediction. For example if Y=0.72, then it is 72% probable (quite likely) that the observed employee will leave.
4. Two approaches we covered:
   - Logistic Regression
   - Decision Tree Classification

Introduction to Classification (cont’d)

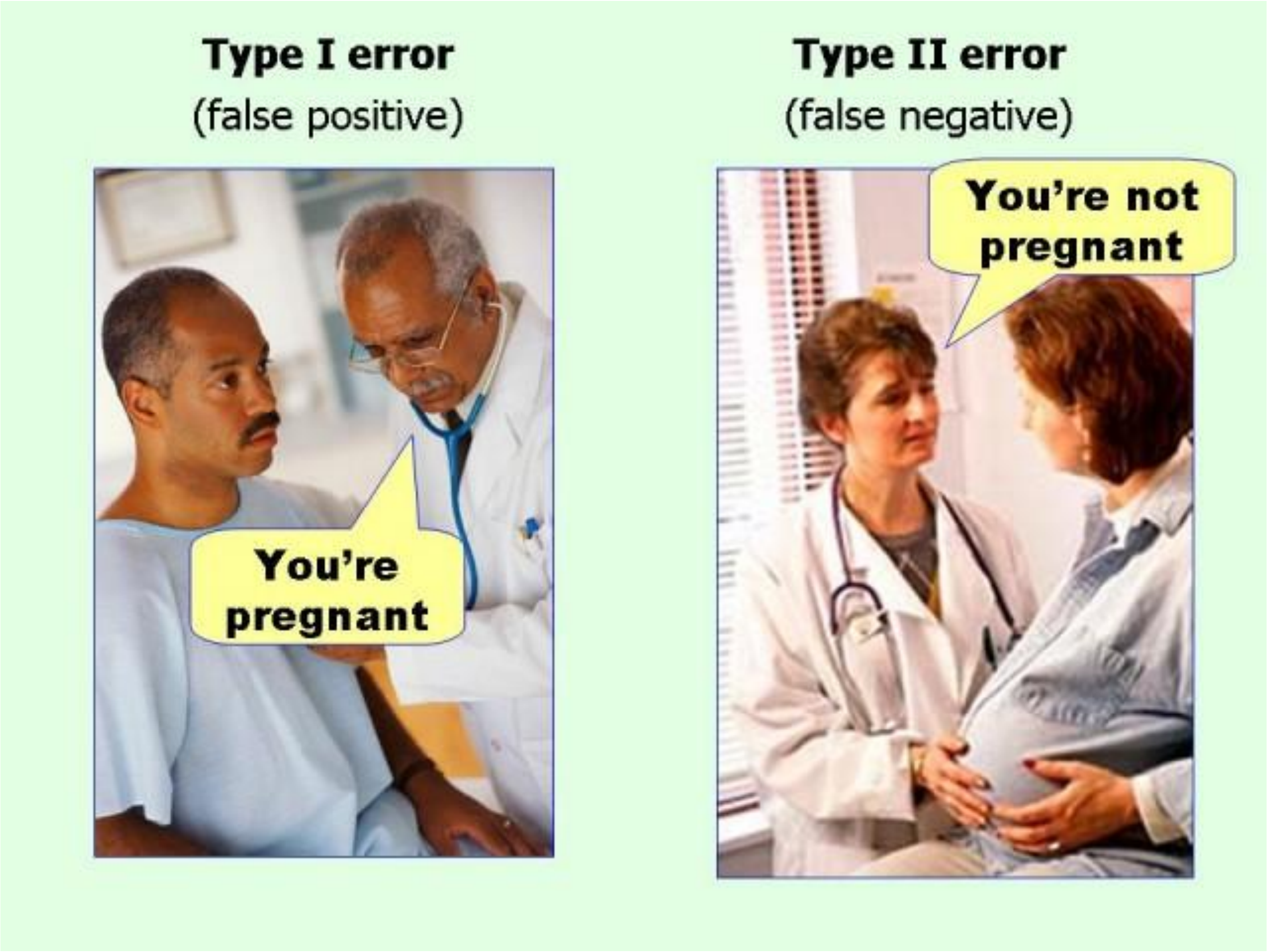
1. Logistic Regression
   - Is simple and fast
   - Is linear
   - Provides probabilities
   - **Is fully interpretable**
   - Is parametric
2. Decision Tree Classification
   - Is simple and fast
   - **Is nonlinear**
   - Does not provide probabilities
   - Is interpretable but not fully (e.g. does not provide p-values)
   - Is nonparametric

Model Evaluation

Prediction accuracy and errors

1. Prediction accuracy is the typical measure used to evaluate classification models.
2. It simply shows the % of correct predictions.
3. The wrong predictions are known as misclassification.
4. There are 2 types of prediction (misclassification) errors:
    - Type I error (false positive): predicting 1(+), while the true value is 0(-)
        - *(e.g. predicting that an employee will leave, while he does not plan to)*
    - Type II error (false negative): predicting 0(-), while the true value is 1(+)
        - *(e.g. predicting that an employee will stay, and the latter leaves)*

Prediction Errors



Confusion Matrix

Confusion Matrix		Reality	
		0	1
Predicted	0	TN	FN
	1	FP	TP

Model evaluation metrics

1. Accuracy alone is enough only for completely balanced datasets.
   - For example, on a dataset where only 5% are 1s (positives), a simple model saying “all are 0s” will produce 95% accuracy but will not be useful at all.
2. If target is to predict positives (e.g. employees who leave), focus on FN:
   - Concentrate on Sensitivity = Recall = True Positive Rate (TPR)
   - Sensitivity (Recall/TPR) = TP/(TP+FN)
   - Sensitivity (Recall/TPR) = % of positives, which were predicted positive
3. If target is to predict negatives (e.g. employees who stay), focus on FP:
   - Concentrate on Specificity = True Negative Rate (TPR)
   - Specificity (TNR) = TN/(TN+FP)
   - Specificity (TNR) = % of negatives, correctly predicted as such

Model Selection

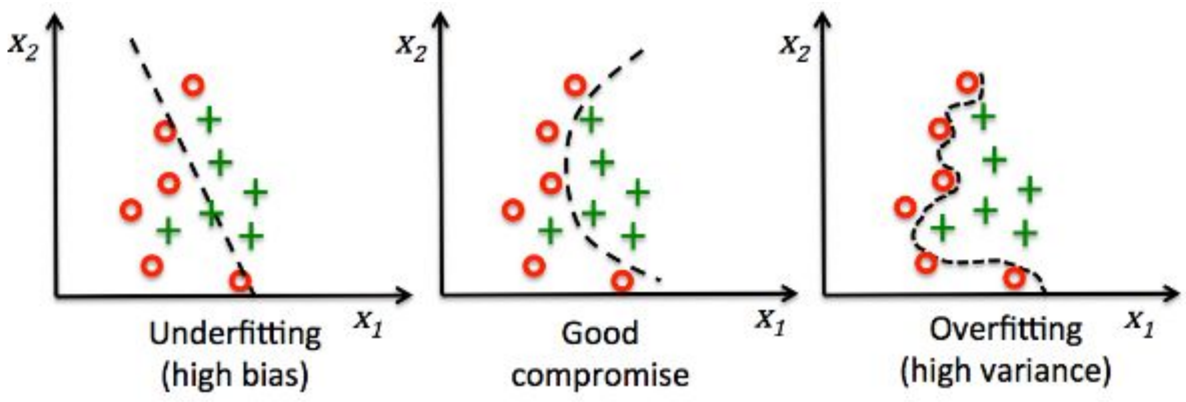
Bias-Variance tradeoff

When making a prediction, we may have 2 objectives:

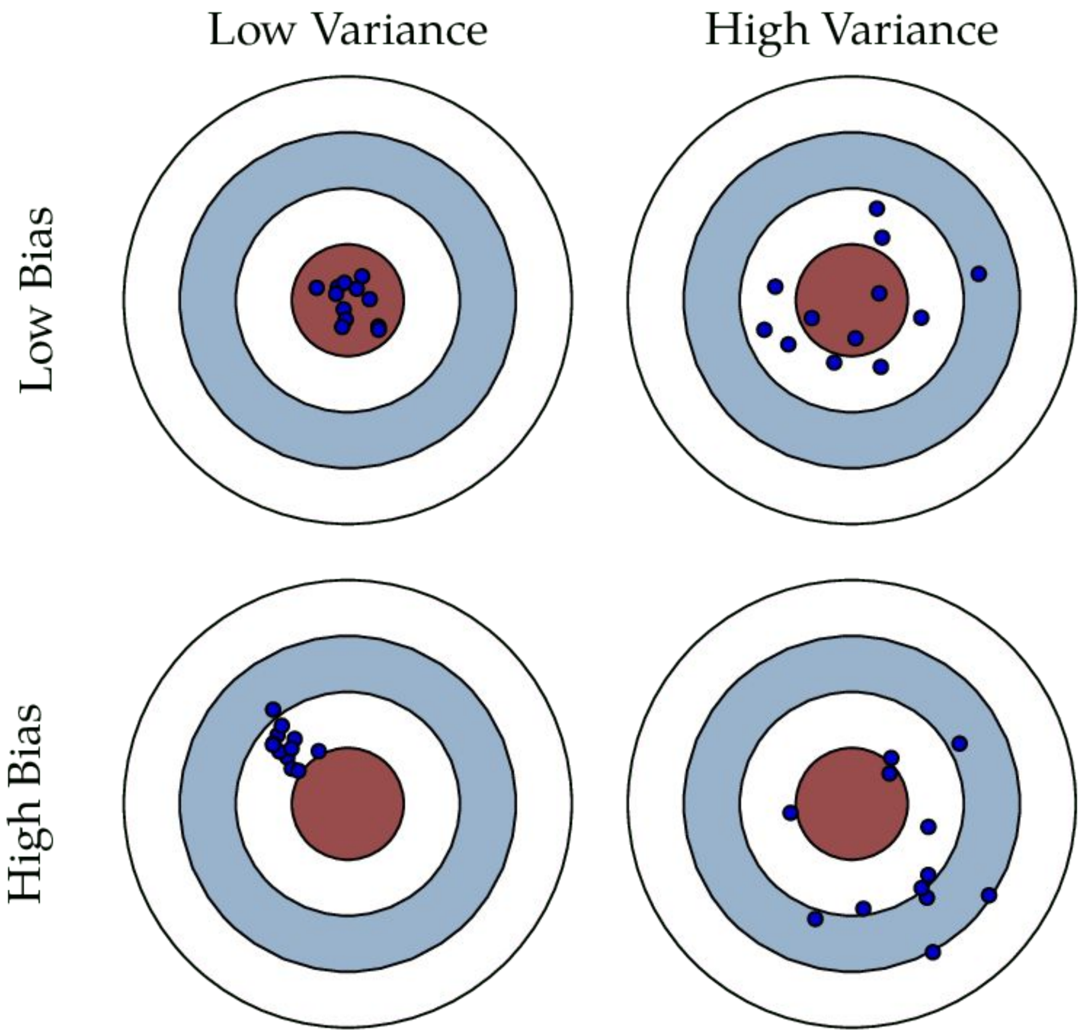
- to get as accurate model as possible (low bias),
- to get as consistent/generalizable model as possible (low variance).

Problem!

- this is not always possible



Bias-Variance tradeoff (cont'd)



Bias-Variance tradeoff (cont’d)

1. The high variance problem is known as overfitting.
2. Overfitting decreases the model generalizability (i.e. model is not useful for external data).
3. To learn about overfitting, the (probably) best solution is train-test split.
4. Develop the model on train, and calculate its accuracy measures on train and test.
5. If they are close, then no overfitting.
6. If 2 measures are different, then you probably have overfitting.

How to fight overfitting?

Develop better models by tuning hyperparameters.

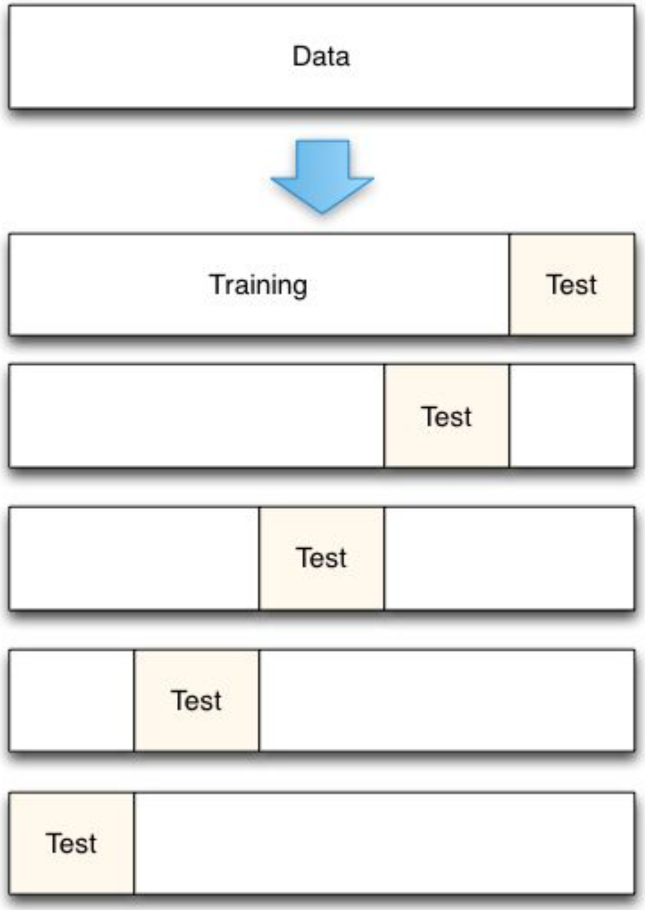
- Either manually put different values for hyperparameters, and choose the best ones.
- Or do the same thing automatically using an algorithm called GridSearch.
- For example: maximum depth of the decision tree to grow.

Problem with Hyperparameter tuning?

- To tune hyperparameters and avoid overfitting the train set, one should calculate accuracy on the test.
- Yet, it is possible, that one starts to overfit train set instead.
- Solution? Evaluate your model on different test sets, not only one.

Cross-Validation

1. Train-test split helps to fight overfitting on the Training data.
2. What if we overfit test data?
3. Solution: test on different Test sets (known as Cross-Validation).
4. Example: 5 fold (5 component) cross validation (on the right).



Thank you