Open | DSE

Introduction to Data Science with Python
Lecture 2: Supervised Learning

Vladimir Osin

Data Scientist/Engineer

Signify Research (formerly known as Philips Lighting)

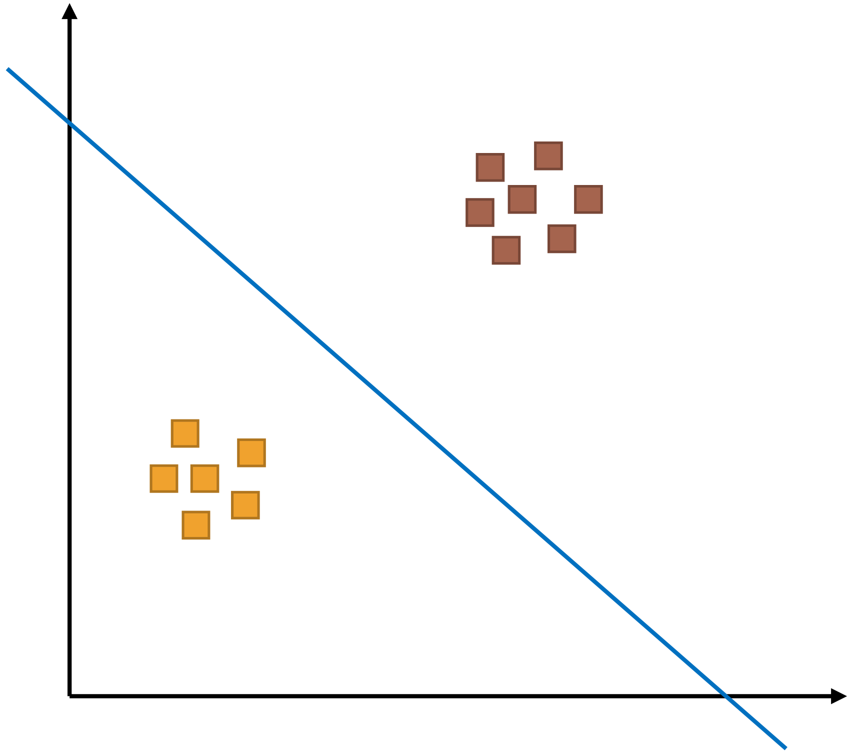
Outline

- Supervised Learning
  - Regression and Classification
    - Linear Regression (Ridge, Lasso)
    - Logistic Regression
    - CART (Classification and regression trees)
    - k-nearest neighbors algorithm
- Evaluation metrics
  - Confusion Matrix
  - Accuracy
  - Precision, Recall
  - F-beta Score
- Concepts
  - Overfitting/Underfitting
  - Cross validation

Supervised Learning

- The core idea of supervised learning is using given set of points that associated to set of outcomes to build a classifier that learns how to **predict outputs from inputs**.
- Type of predictions: continuous (regression) and class (classification)

- Linear models
- Non-parametric approaches
  - Tree-based models
  - Nearest Neighbours methods



Linear Regression

The main assumption is that predicted value is expected to be a **linear combination** of the input variables.

- Simple Linear Regression
    - Minimization the sum of squared residuals
    - Highly sensitive to random errors

$$\min_{w} ||Xw - y||_2{}^2$$

- Ridge regression
    - Introducing L2 penalty for weights
    - Alpha parameters helps to control shrinkage

$$\min_{w} ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$$

- Lasso regression
    - Introducing L1 penalty for weights
    - Effectively reducing the number of variables

$$\min_{w} \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1$$

Logistic Regression

- Used for **classification,** actually non-linearity over a linear classifier
- Predict the probability of categorical dependent variable
- Types: binary(sigmoid activation), multi-class(softmax activation)
- Loss function: cross-entropy

$$J(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} H(p_n, q_n) = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1 - y_n)\log(1 - \hat{y}_n)\right]$$

Trees (CART)

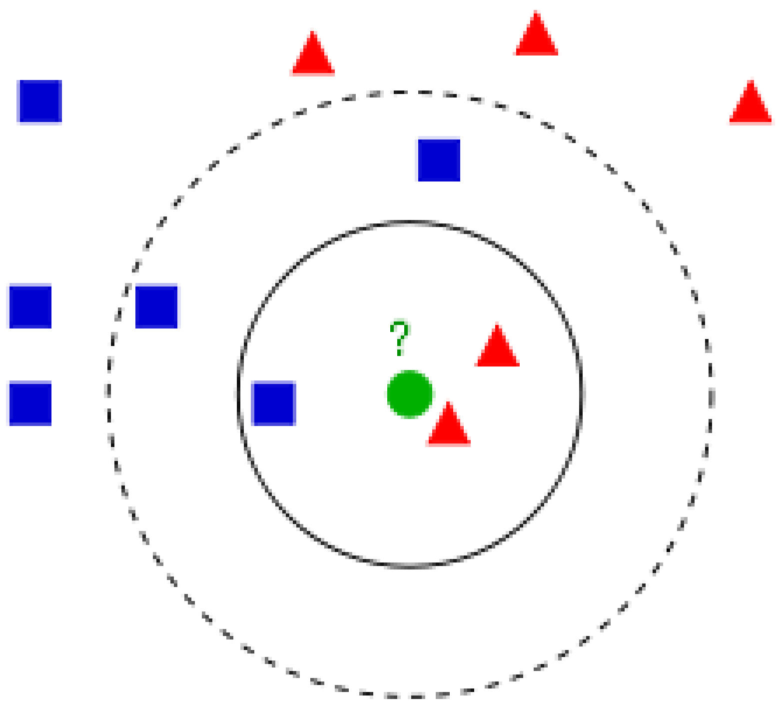
- Decision Tree 
  - Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

- Random Forest 
  - Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

- Gradient-Boosted Tree 
  - Classifier is trained on data, taking into account the previous classifiers' success. After each training step, the weights are redistributed. Misclassified data increases its weights to emphasize the most difficult cases. In this way, subsequent learners will focus on them during their training.

Open | DSE

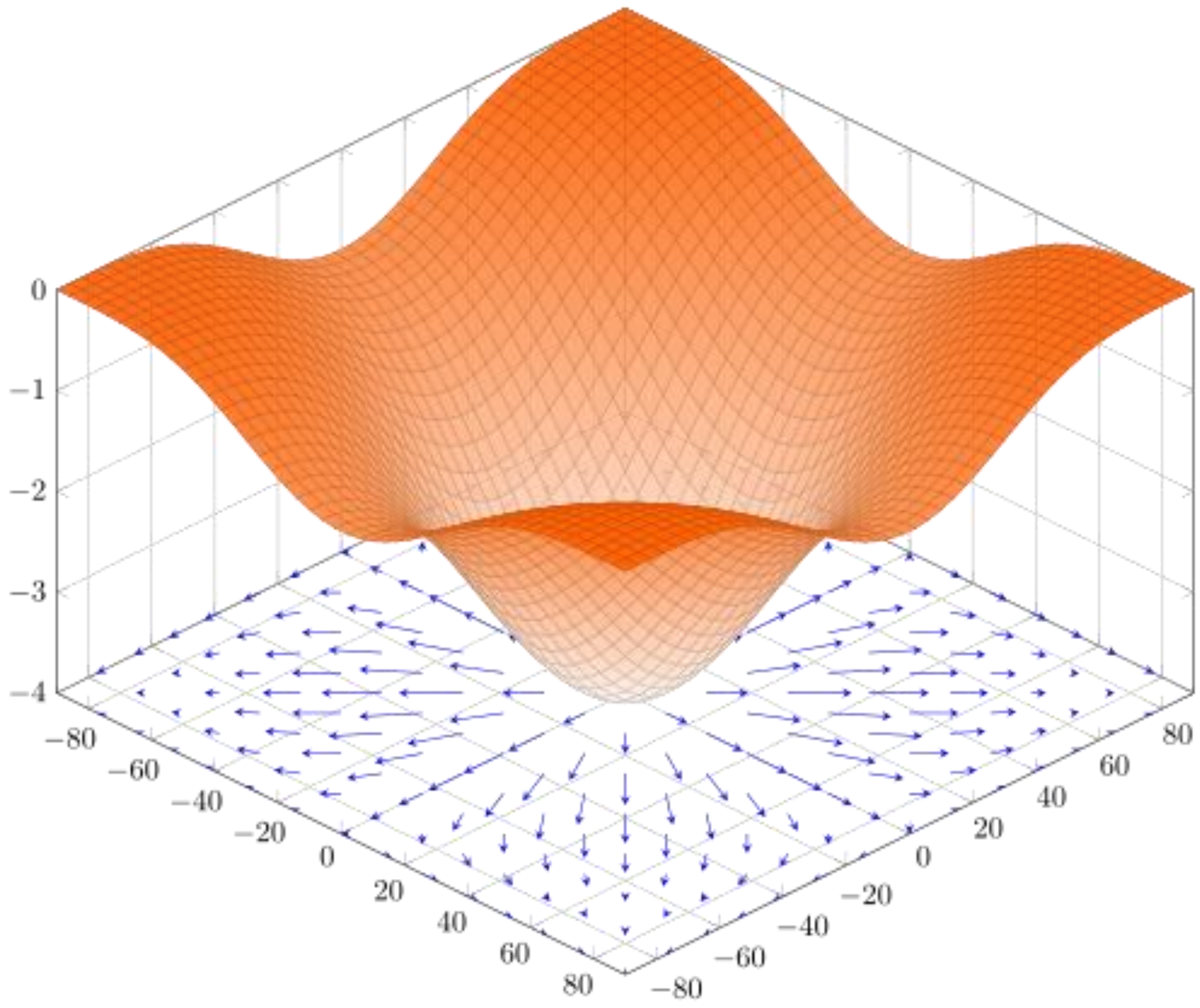
Nearest-Neighbor Methods



- Idea: find some amount of closest points (with classes that we know) and then average their responses
  - Metric: Euclidian Distance
  - Classification: Majority Vote
  - Regression: Average of values
- Sensitive to local structure of the data

Stochastic Gradient Descent

- Gradient points in the direction of the greatest rate of increase of the function
- Descending the gradient = moves to anti-gradient direction



Evaluation Metrics (Classification)

		Predicted class +	Predicted class −
Actual class	+	**TP** True Positives	**FN** False Negatives Type II error
Actual class	−	**FP** False Positives Type I error	**TN** True Negatives

Evaluation Metrics (Classification)

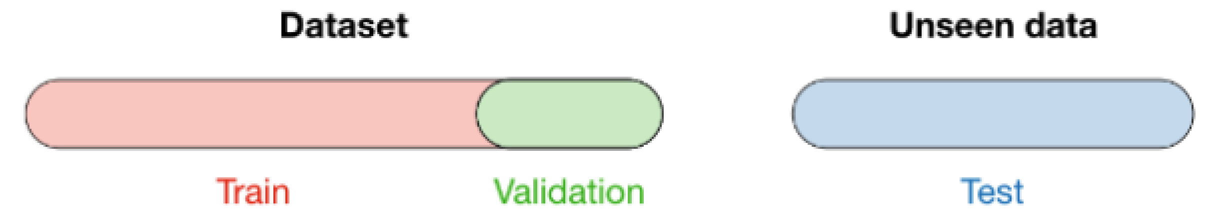
Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN + FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

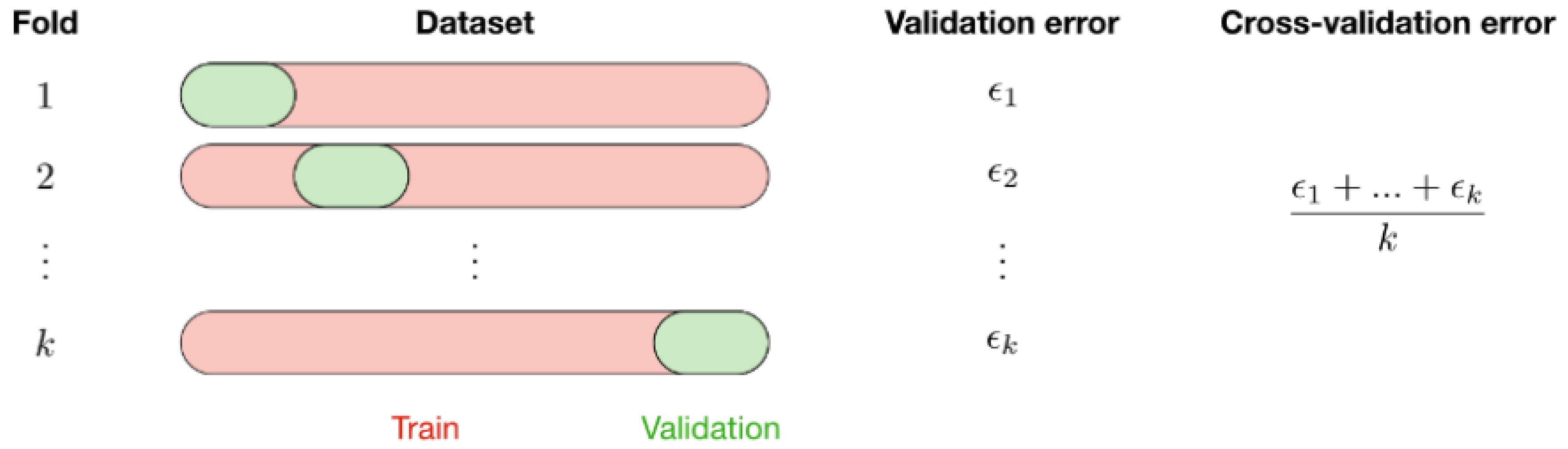
Evaluation Metrics (Regression)

- Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error
- Root Mean Squared Logarithmic Error

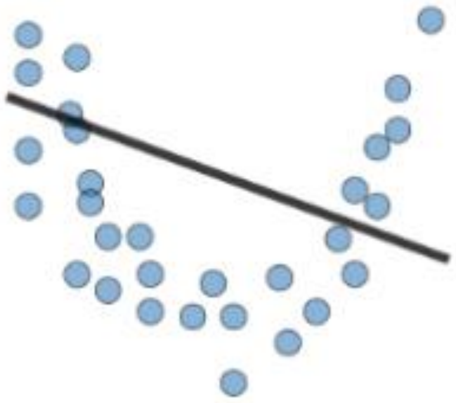
- Coefficient of determination (R2 score)

Cross-validation





Bias/variance tradeoff

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			

Open | DSE

Practice

- http://scikit-learn.org/stable/user_guide.html
- http://scikit-learn.org/stable/supervised_learning.html
- http://scikit-learn.org/stable/auto_examples/index.html

Assignment 2

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Getting Started · Ongoing · tutorial, tabular data, binary classification

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Getting Started · Ongoing · tabular data, regression