Data Mining Review

CS 584: Big Data Analytics

What is Data Mining?

- Definition (Fayyad, Piatetsky-Shapiro and Smyth, 96) "Knowledge Discovering in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."
- Patterns can be anything from associations, groupings, trends, anomalies, etc.
- "Extraction of useful information from large data sets"

Types of Models

Consider a large collection of fruits with attributes or characteristics such as weight, volume, color, shape…

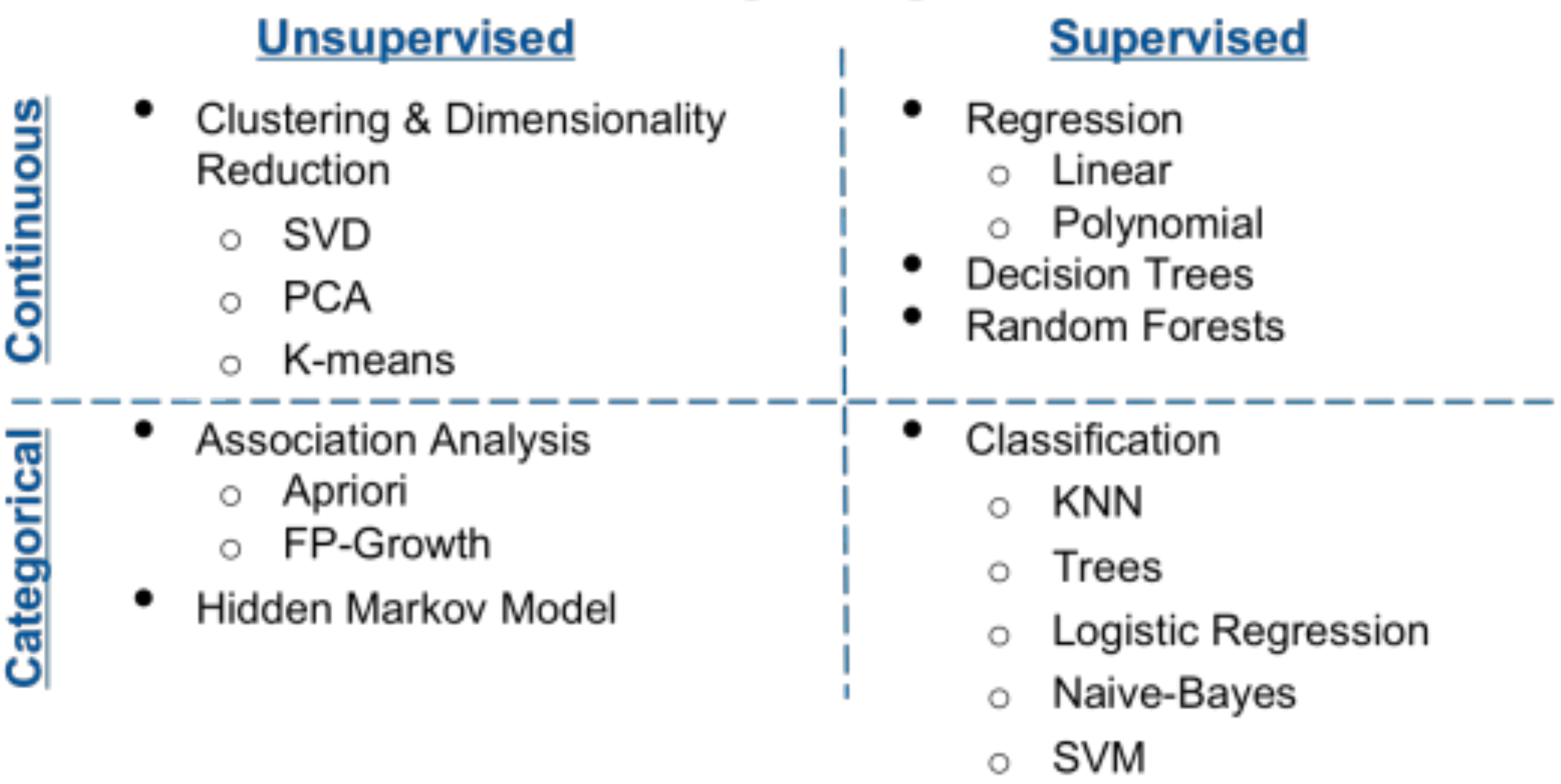
- **Regression**: predicting weight based on other attributes
- **Classification**: what type of fruit is it?
- **Clustering**: how many different types of fruits are there?
- **Anomaly detection**: is this fruit different than its other types?
- …

Adapted from Joydeep Ghosh's course in Data Mining

Degrees of Supervision in Learning Algorithms

- **Unsupervised**
  - Model is not provided correct results during the training
  - Cluster based on input data and statistical properties
- **Supervised**
  - Training data includes input and desired results
- **Semi-supervised**
  - Tasks that make use of unlabeled data for training with a small amount of labeled data

Data Mining Algorithms

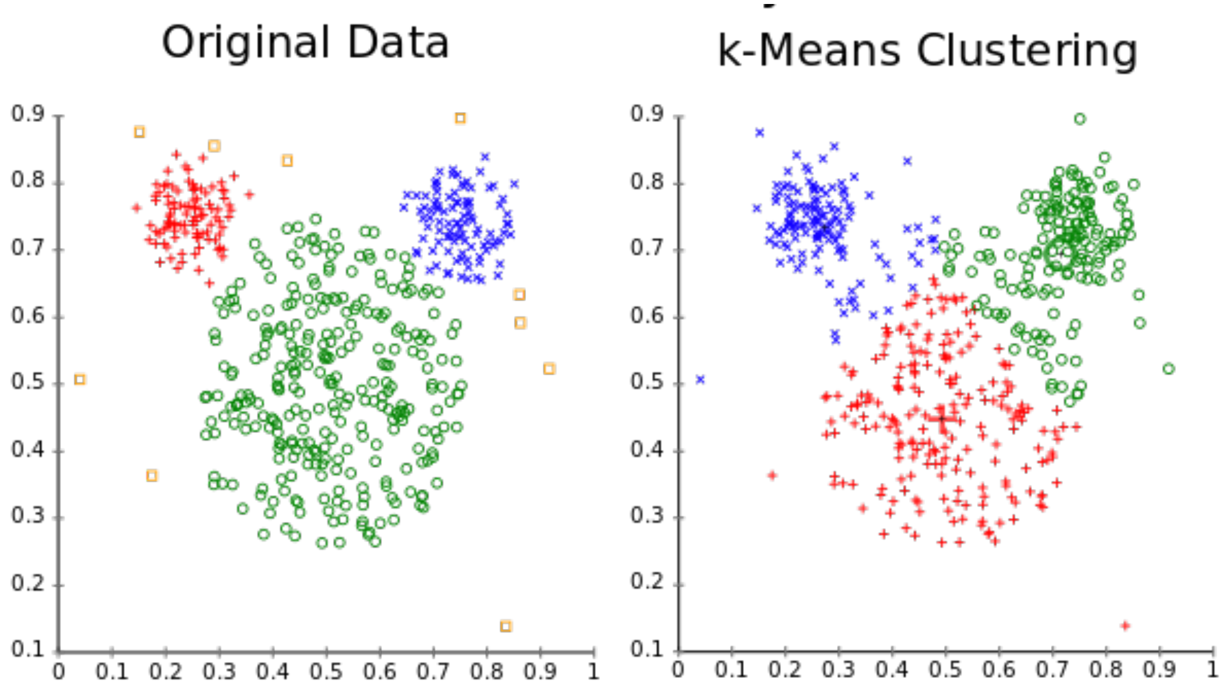


	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

https://nyghtowlblog.files.wordpress.com/2014/04/ml_algorithms.png?w=535&h=311

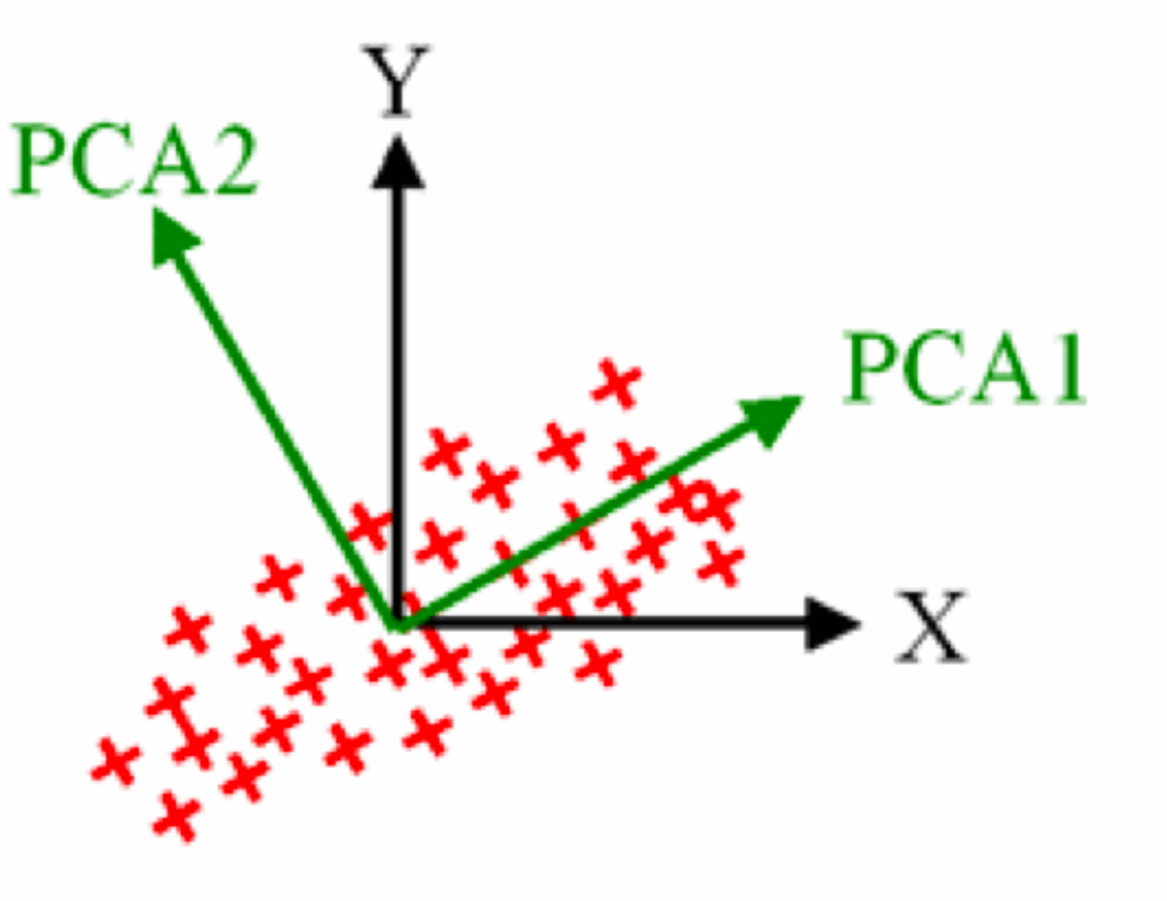
Clustering: K-Means

- Partition the data into k clusters based on their features
- Each cluster is represented by its centroid, defined as the center of the points in each cluster
- Each point is assigned to the cluster whose center is nearest



Clustering: PCA and SVD

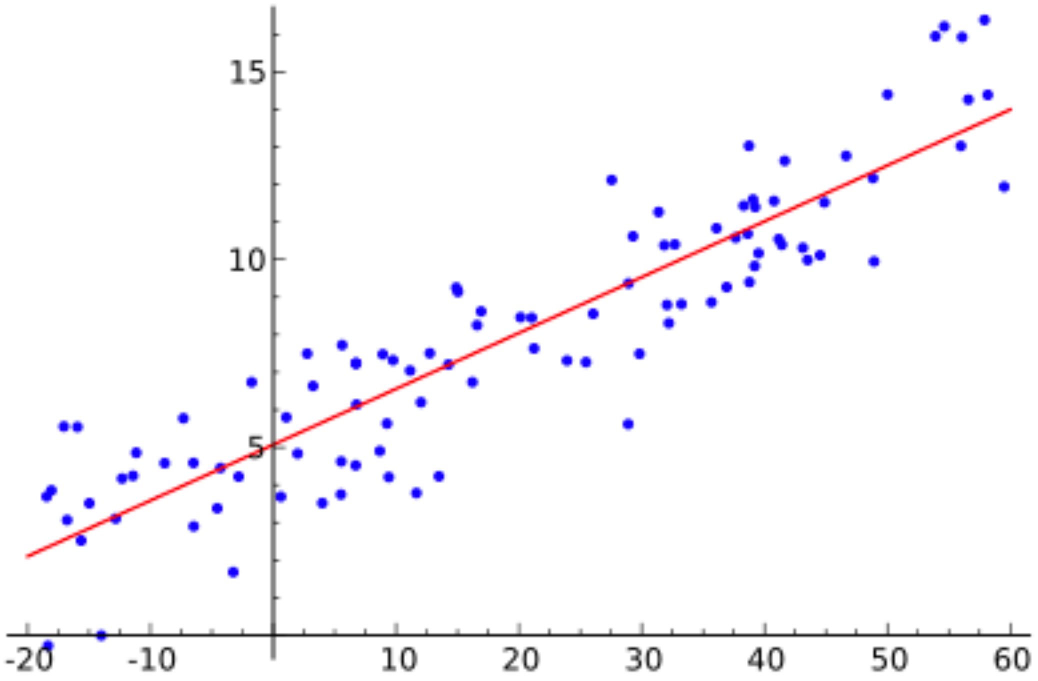
- Linear projection of high dimensional data into a lower dimensional subspace
  - Variance retained is maximized
  - Least square reconstruction error is minimized
- Baseline matrix factorization method: best possible matrix approximation given number of components

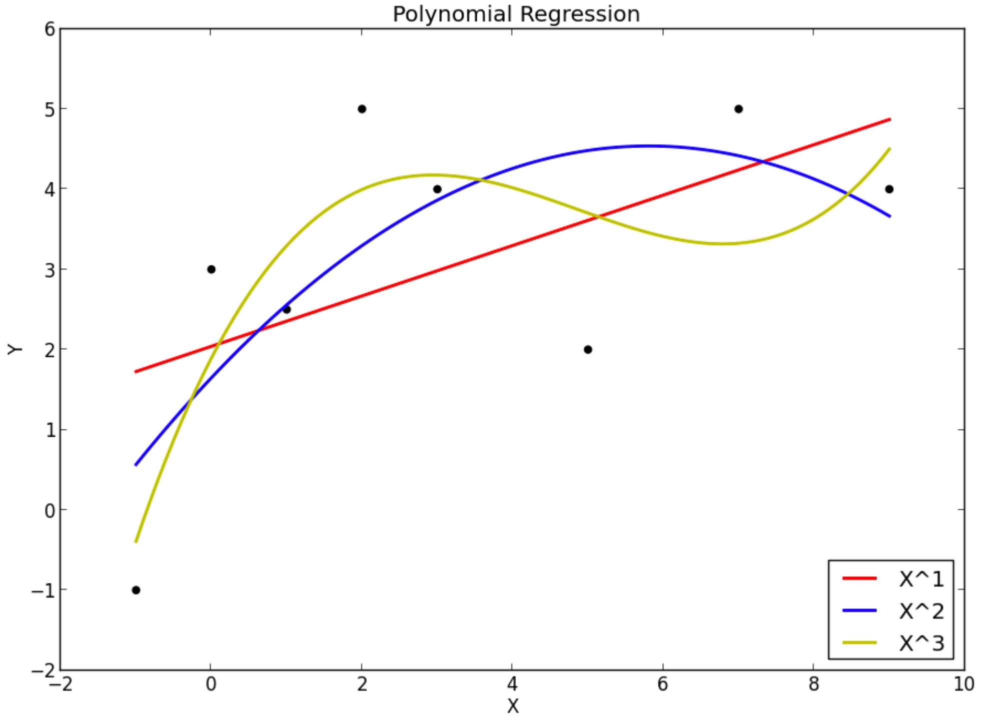


https://prateekvjoshi.files.wordpress.com/2014/10/2-pca.png

Regression

- Predict new values based on the past
- Compute new values of a dependent variable based on values of one or more measured attributes





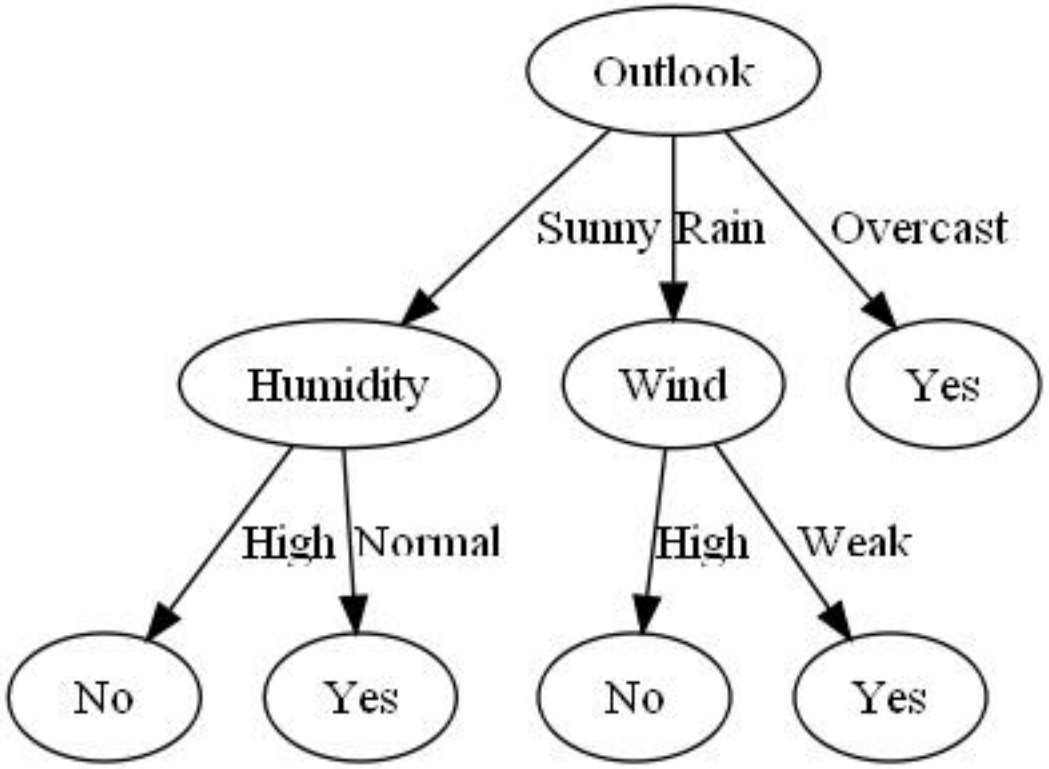
Classification: Decision Trees

- A simple set of rules to classify your data

- Splitting criteria determines the rules that will be derived from the data

- Non-parametric because there are no assumptions about the distribution of the variables of each class

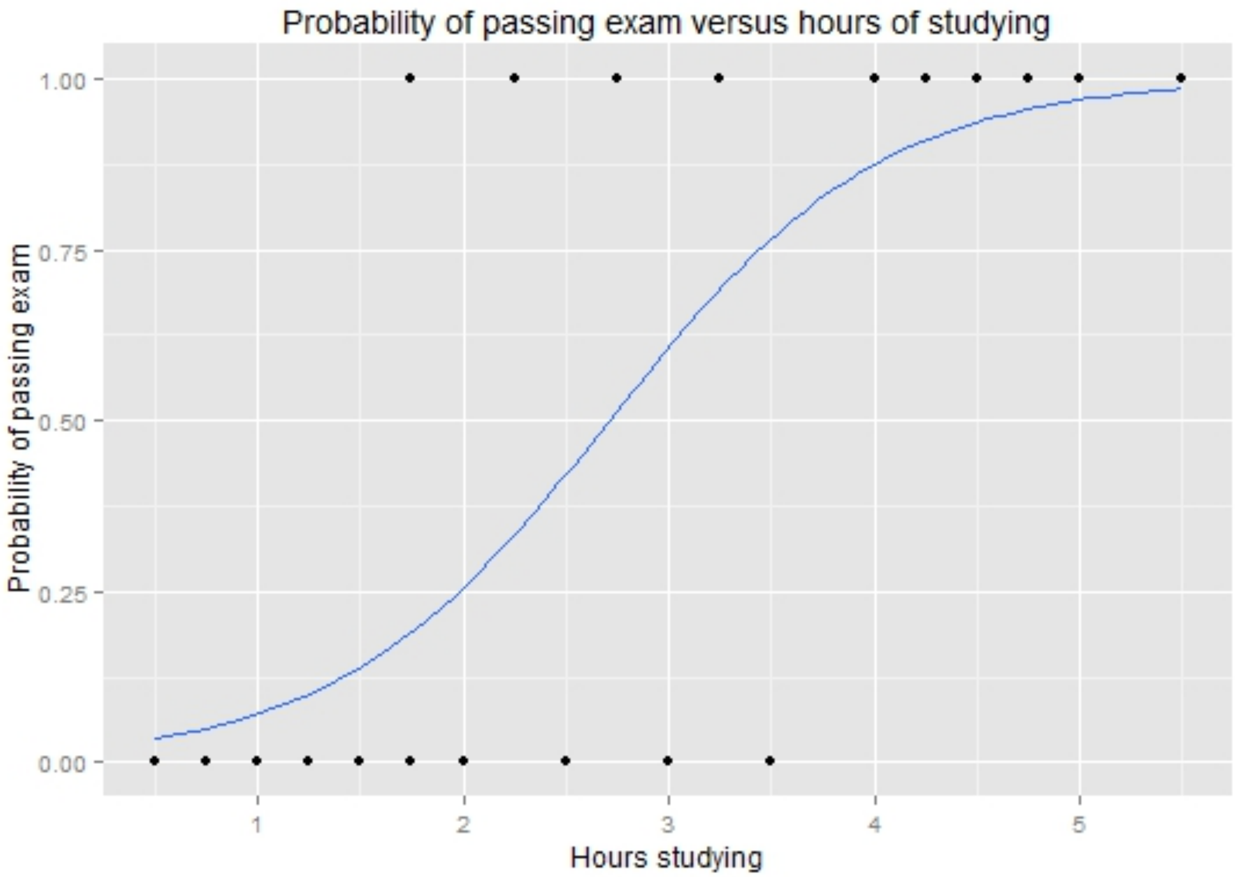
Should i play tennis?



http://blogs.msdn.com/blogfiles/chrsmith/WindowsLiveWriter/AwesomeFDecisionTreesPartI_131F5/image_3.png

Classification: Logistic Regression

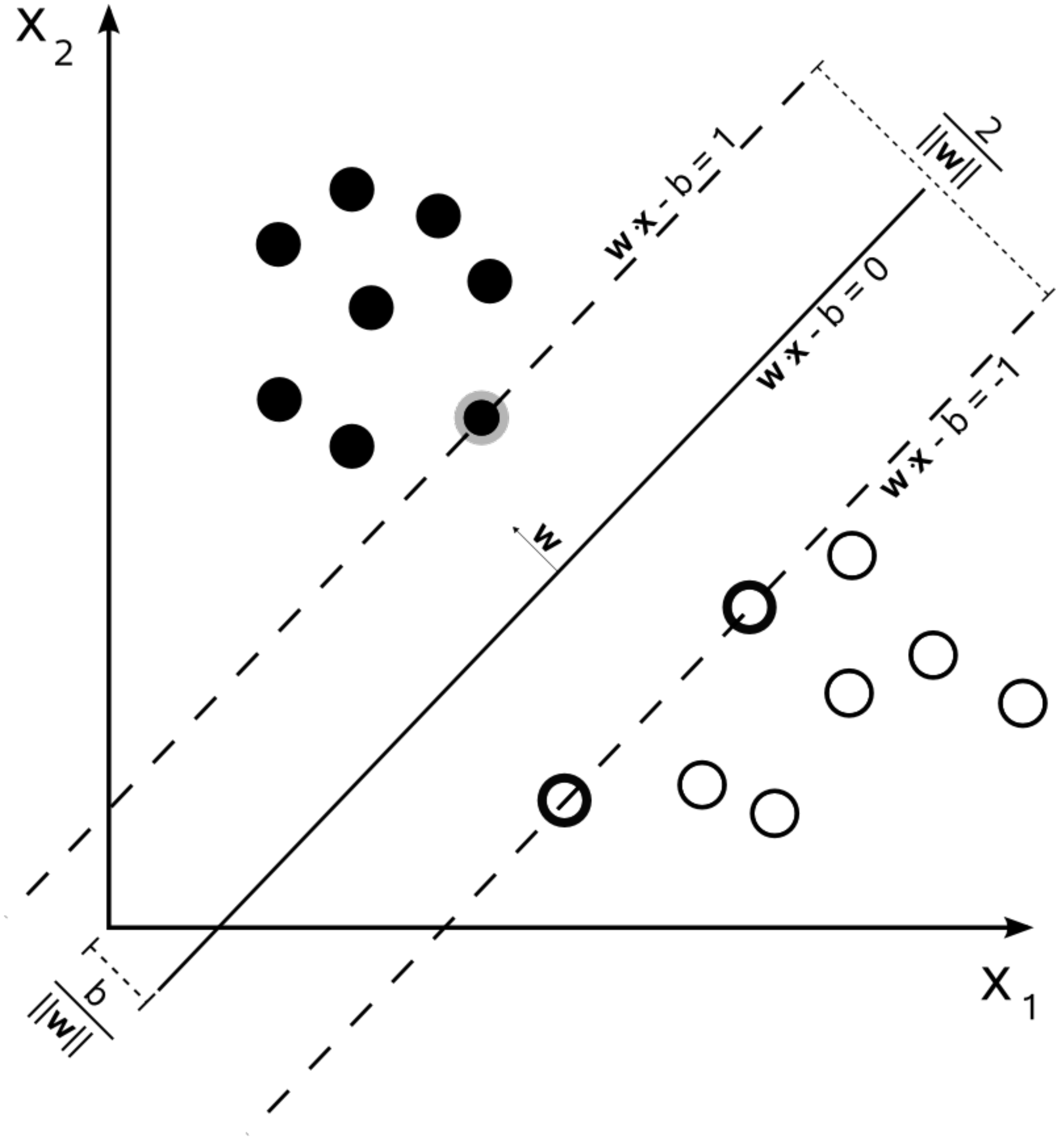
- Regression model with the dependent variable is categorial (known as a generalized linear model)
- Logistic function to transform linear model
- Produces log-odds ratio as a linear function of predictors



https://en.wikipedia.org/wiki/Logistic_regression

Classification: SVM

- Use optimal hyperplane in a suitable feature space for classification
- Flexible représentation of class boundaries
- Allows nonlinear in original features using the “kernel trick”

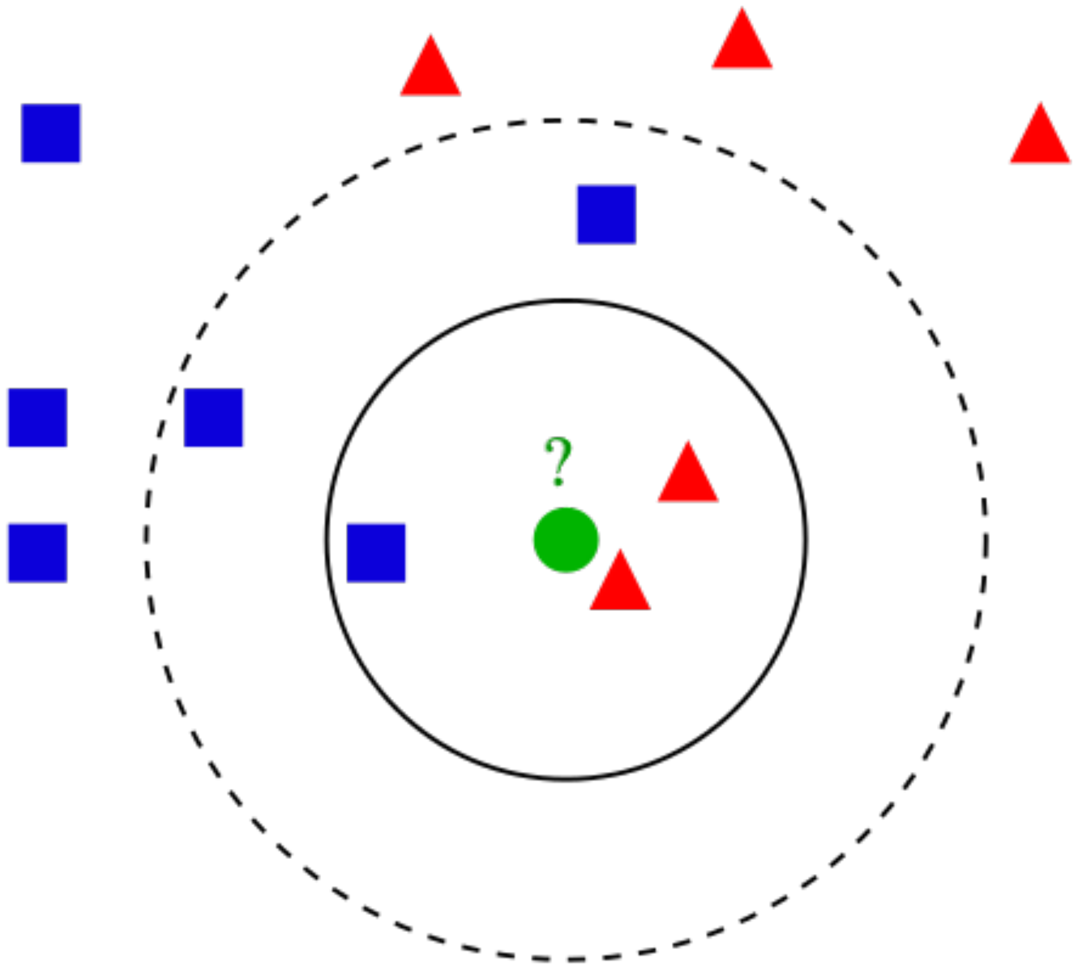

x_2
$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w}\cdot\mathbf{x} - b = 1$
$\mathbf{w}\cdot\mathbf{x} - b = 0$
$\mathbf{w}\cdot\mathbf{x} - b = -1$
$\mathbf{w}$
$\frac{b}{\|\mathbf{w}\|}$
x_1

Classification: k-Nearest Neighbors (KNN)

- Example of instance-based learning, or lazy learning

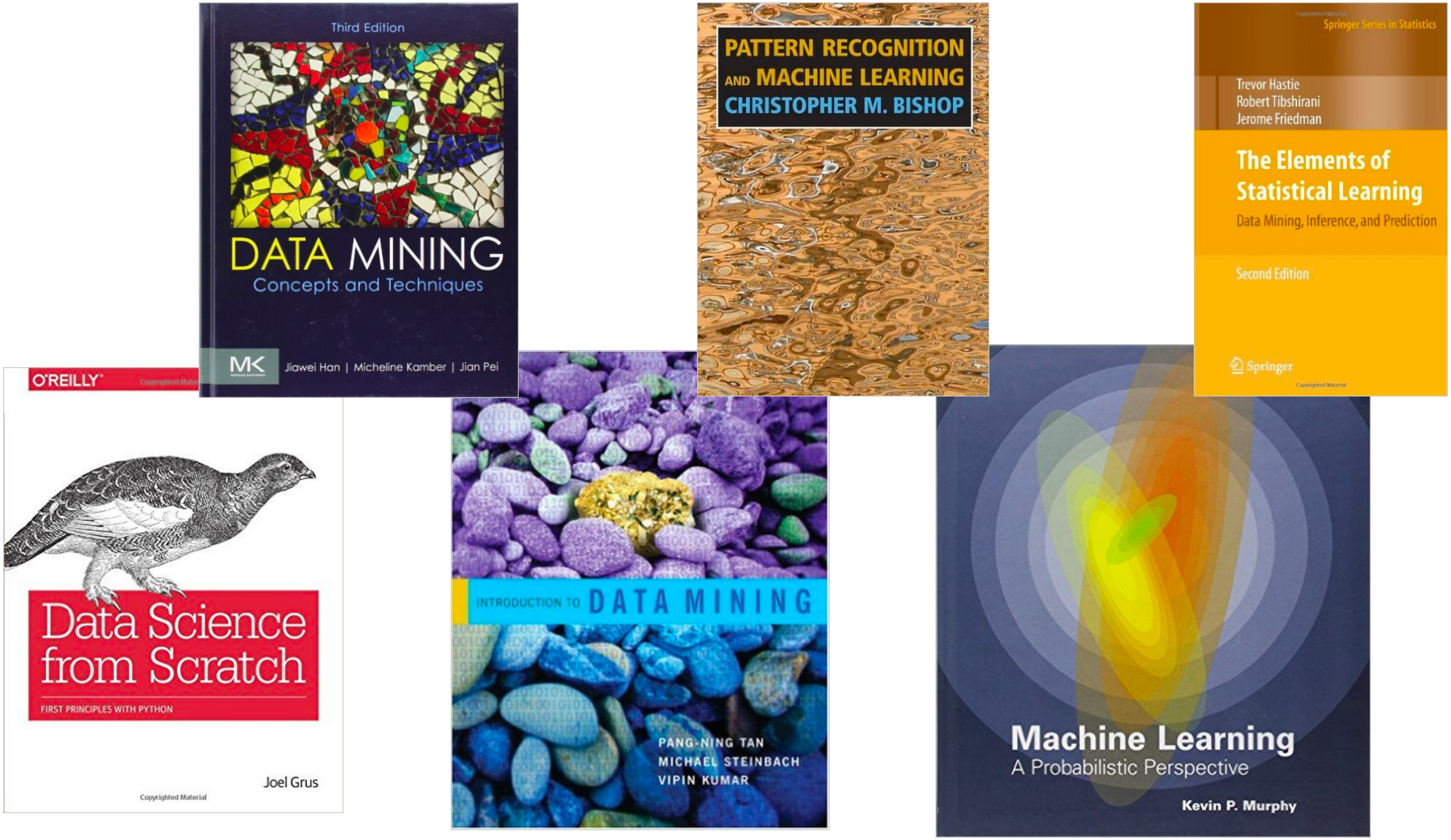
- Find the k nearest neighbors based on distance metric and classify by assigning the label which is most frequent in the k samples

- Smaller k - local, larger k - global



https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Data Mining Books


Third Edition
DATA MINING
Concepts and Techniques
MK
Jiawei Han | Micheline Kamber | Jian Pei
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer
O'REILLY
Data Science from Scratch
FIRST PRINCIPLES WITH PYTHON
Joel Grus
INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR
Machine Learning
A Probabilistic Perspective
Kevin P. Murphy

Data Mining & Python Resources

- Continually updated data science Python notebooks
https://github.com/donnemartin/data-science-ipython-notebooks

- Applied Predictive Modeling Book & Python
http://nbviewer.jupyter.org/github/leig/Applied-Predictive-Modeling-with-Python/tree/master/notebooks/

- Data Science Resources
https://github.com/jonathan-bower/DataScienceResources

- Interesting iPython Notebooks
https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks