Machine Learning: Gentle Introduction

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 790-784, Korea
seungjin@postech.ac.kr
Textbooks, References, and Web

(a) Murphy

(b) Bishop

(c) Duda et al.

http://mlg.postech.ac.kr/~seungjin/course/ml/ml.html
Machine Learning: A scientific discipline that is concerned with the design and development of algorithms that allow computers to learn from empirical data (sensor data or database) and to make predictions.

Descriptive or Predictive Learning
Probabilistic Models

Data makers

- Data $X = [x_1, x_2, \ldots, x_N]$ are generated from an unknown underlying distribution $p(x)$.

Models

- Assume a model distribution $p(x_1, \ldots, x_N|\theta)$ that is controlled by parameters $\theta$.

Consider a (directed or undirected) graph where nodes are associated with random variables and edges represent statistical dependencies between variables.

- undirected graph
- directed graph
- factor graph
Maximum Likelihood vs Bayesian

Given a set of data points, \( X = [x_1, \ldots, x_N] \), density estimation involves estimating a distribution from which observed data points \( x_i \) are drawn.

**Maximum Likelihood**
- Model \( x_i | \theta \sim p(\cdot | \theta) \).
- Find MLE:
  \[
  \theta_{ML} = \arg \max_{\theta} \log p(X | \theta) \\
  = \arg \max_{\theta} \sum_{i=1}^{N} \log p(x_i | \theta) .
  \]
- Prediction is done by
  \[
  p(x_\ast | \theta_{ML}) .
  \]

**Bayesian**
- Model \( x_i | \theta \sim p(\cdot | \theta) \).
- Prior over parameters: \( p(\theta) \).
- Posterior over parameters
  \[
  p(\theta | X) = \frac{p(X | \theta)p(\theta)}{p(X)} .
  \]
- Prediction is done by
  \[
  p(x_\ast | X) = \int p(x_\ast | \theta)p(\theta | X) d\theta .
  \]
Unsupervised vs Supervised

Unsupervised (descriptive) learning
Given a set of unlabeled examples, \( \mathcal{D} = \{x_i\}_{i=1}^{N} \), learn a fruitful representation (or statistical regularities) of the data.

- Discovering clusters
- Discovering latent factors (dimensionality reduction)
- Discovering graph structure
- Matrix completion

Supervised (predictive) learning
Given a set of labeled examples, \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \), learn a mapping \( f: \mathcal{X} \mapsto \mathcal{Y} \), such that given an unseen example \( x_* \), associated output \( y_* \) is predicted.

- Classification
  - The desired outputs \( y_i \) are discrete class labels.
  - The goal is to classify new inputs correctly.
- Regression
  - The desired outputs \( y_i \) are continuous valued.
  - The goal is to predict the output accurately for new inputs.
- Ordinal regression
Clustering

- Mixture of Gaussians
- \( k \)-means
- Hierarchical clustering
- Spectral clustering

(a) MoG

(b) \( k \)-means

(c) Hierarchical

(d) Spectral
Latent Class vs Latent Feature Models

\[ p(x_i | z_i) = \prod_{k=1}^{K} \mathcal{N}(x_i | \mu_k, \Sigma_k)^{z_{ki}}. \]

\[ p(x_t | s_t) = \mathcal{N}(x_t | As_t, \Sigma). \]
Blue crescent should be \( y = 1 \) since all blue shapes are labeled 1 in the training set.

Yellow circle is harder to classify since some yellow things are labeled \( y = 1 \) and some are labeled \( y = 0 \), and some circles are labeled \( y = 1 \) and some \( y = 0 \).

The correct label for the blue arrow is unclear.

**Probabilistic prediction:** Compute the best guess

\[
\hat{y}_* = \arg \max_c p(y_* = c | x_*, X).
\]
Regression

Regression model: \( y = f(x) + \epsilon. \)

- **Linear regression**: \( f(x) = \sum_{j=1}^{M} w_j \phi_j(x) + w_0. \)
- **Kernel regression**: \( f(x) = \sum_{i=1}^{N} w_i k(x, x_i) + w_0. \)
Course Outline

http://mlg.postech.ac.kr/~seungjin/ml.html

- Density estimation
- Clustering
  - $k$-means clustering
  - Mixture of Gaussians
  - Nonnegative matrix factorization
  - Spectral clustering
- Latent variable models
  - EM optimization
  - Factor analysis and PCA
  - Mixture of factor analyzers
  - Kernel PCA
  - Restricted Boltzmann machines
- Regression
  - Linear models for regression
  - Logistic regression
  - Mixture of experts
- Classification
  - Bayes decision theory
  - Linear discriminant analysis
  - Deep learning and Neural networks
  - Support vector machines
  - Hidden Markov models