Summary of Clustering and Linear Models

CS 540

Yingyu Liang

Hierarchical Clustering

Hierarchical Clustering

- Input: data set $\{x_i\}$, a distance function between clusters
- Output: a hierarchy on the data points

1. Initialize each point as an individual cluster
2. Repeat until only one cluster remains:
   - Find the closest pair of clusters
   - Merge the pair into one cluster
3. Output the tree where leaves are the data points, and the internal nodes correspond to merges performed

Hierarchical Clustering

- Single-linkage: the shortest distance from any member of one cluster to any member of the other cluster. Formula:
$$d(A,B) = \min_{x \in A, y \in B} d(x,y)$$
- Complete-linkage: the greatest distance from any member of one cluster to any member of the other cluster
$$d(A,B) = \max_{x \in A, y \in B} d(x,y)$$
- Average-linkage: the average distance from any member of one cluster to any member of the other cluster

$$d(A,B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x,y)$$

K-means Clustering

K-means Clustering: Objective

- Input: data set $\{x_i\}$ where each data point is a numeric feature vector in R^d, the number of clusters k

- Would like to get a clustering with a small distortion

$$\min_{\substack{y(x_1),y(x_2),\ldots,y(x_n)\\ c_1,c_2,\ldots,c_k}} \sum_{x\in\{x_i\}} \left\|x - c_{y(x)}\right\|_2^2$$

K-means Clustering: Deriving the Algorithm

- If fix centers:

$$\min_{y(x_1),y(x_2),\ldots,y(x_n)} \sum_{x\in\{x_i\}} \left\|x - c_{y(x)}\right\|_2^2$$

Only need to assign each point to its closest center

- If fix assignments:

$$\min_{c_1,c_2,\ldots,c_k} \sum_{x\in\{x_i\}} \left\|x - c_{y(x)}\right\|_2^2$$

Only need to set each center to be the average of points in the cluster

K-means Clustering: Algorithm

- Input: data set $\{x_i \in R^d\}$, number of clusters k
- Output: k clusters and their centers

1. Initialize k cluster centers
2. Repeat until convergence:
   - Assign each point to its closest center
   - Update each center to be the average of data points in the cluster
3. Output the k clusters and their centers

Linear Regression

Linear Regression: Model

- Input: data set $\{(x_i, y_i)\}$ where $x_i \in R^{p+1}, y_i \in R$
- Model: $y = f(x) = \beta^T x$, where $\beta \in R^{p+1}$

- Assumption: there is ground truth β^* and the label is given by

$$y = (\beta^*)^T x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

Linear Regression: Deriving the Objective

- Maximum Likelihood Estimate (MLE) leads to Ordinary Least Squares (OLS)

$$\hat{\beta} = \text{argmin}_\beta \, \|\boldsymbol{y} - X\beta\|_2^2$$

where $X \in R^{n\times(p+1)}$ be a matrix where the i-th row is x_i
and $\boldsymbol{y} \in R^n$ be a vector where the i-th entry is y_i

- Maximum A Posteriori (MAP) leads to Ridge Regression

$$\hat{\beta} = \text{argmin}_\beta \, \|\boldsymbol{y} - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

where $\lambda > 0$ is the regularization coefficient. $\lambda = 0$ leads to OLS.

Linear Regression: Solving the Optimization

- Ridge Regression

$$\hat{\beta} = \mathrm{argmin}_{\beta} \, \|\boldsymbol{y} - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

- Convex optimization
- Setting gradient to 0:

$$-2X^T\boldsymbol{y} + 2X^TX\beta + 2\lambda\beta = 0$$

$$\hat{\beta} = (X^TX + \lambda I)^{-1}X^T\boldsymbol{y}$$

(OLS is the special case with $\lambda = 0$, so need X^TX invertible)

Logistic Regression

Logistic Regression: Model

- Input: data set $\{(x_i, y_i)\}$ where $x_i \in R^{p+1}, y_i \in \{+1, -1\}$
- Model: $p(y = +1|x) = \sigma(\theta^T x)$, where $\theta \in R^{p+1}, \sigma(z) = \frac{1}{1+\exp(-z)}$
  - Can predict label $+1$, if $\sigma(\theta^T x) \geq 0.5$

- Assumption: there is ground truth θ^* and the label is given by

$$p(y = +1|x) = \sigma\left((\theta^*)^T x\right)$$

Logistic Regression: Deriving the Objective

- Maximum Likelihood Estimate (MLE) leads to:

$$\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{i}} \log\left(1 + \exp(-y_i \theta^T x_i)\right)$$

- Maximum A Posteriori (MAP) leads to:

$$\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{i}} \log\left(1 + \exp(-y_i \theta^T x_i)\right) + \frac{\lambda}{2} \|\theta\|_2^2$$

where $\lambda > 0$ is the regularization coefficient. $\lambda = 0$ leads to the MLE objective.

Logistic Regression: Solving the Optimization

- Regularized logistic regression:

$$\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{i}} \log\left(1 + \exp(-y_i \theta^T x_i)\right) + \frac{\lambda}{2} \|\theta\|_2^2$$

- Convex optimization
- But no closed form solution; solve via (stochastic) gradient descent