Model order: M+1

Feature matrix:
N by M+1

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^M \\ 1 & x_2 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^M \end{bmatrix}$$

Linear basis model: $y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j \phi_j(x) = \mathbf{X}\mathbf{w}$

Objective Function, mse: $J(\mathbf{w}) = \frac{1}{2} \|\mathbf{t} - \mathbf{y}\|_2^2$

$\arg_{\mathbf{w}} \min J(\mathbf{w}) : \mathbf{w} \ = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T\mathbf{t}$

Ridge(L2): $= \lambda \sum_{i=0}^{M} w_i^2 \ = \lambda \|\mathbf{w}\|_2^2$

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1} \mathbf{X}^T\mathbf{t}$$

Lasso(L1): $= \lambda \sum_{i=0}^{M} |w_i|$
Will approach 0

ElasticNet: $= \beta\lambda \sum_{i=0}^{M} |w_i| + (1-\beta)\lambda \sum_{i=0}^{M} w_i^2$

$\arg_{\mathbf{w}} \max \exp\left(-J(\mathbf{w})\right)$

MAP(Bayesian): $L^0 = \prod_{i=1}^{N} P(x_i|\mu)P(\mu)$

MLE(Frequentist):: $L^0 = \prod_{i=1}^{N} P(\mu|x_i)$

log-likelihood
$L = ln(L^0)$

Univariate Gaussian: $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$

Laplacian: $\frac{1}{2b} \exp\left\{-\frac{|x-\mu|}{b}\right\}$

Bernoulli: $\mu^x(1-\mu)^{1-x}$ -> $\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Beta: $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$

- constant: $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

$$\alpha^{t+1} = \alpha^t + \sum_{i=1}^{N} x_i$$

$$\beta^{t+1} = \beta^t + \sum_{i=1}^{N} x_i$$

Bernoulli-Beta:(posterior)-(prior) $\mu_{MAP} = \frac{\alpha - 1 + \sum x_i}{N + \alpha + \beta - 2}$

Multivariate Gaussian:

$$\frac{1}{(2\pi)^{N/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}$$

- Σ is the covariance matrix of k
- $|\Sigma| = (\sigma^2)^N$ is its determinant of k
- Σ^{-1} is the inverse of k

Assuming Σ is isotropic:

Sample Avg: $\mu_k = \frac{1}{N_k}\sum_{i=i}^{N_k} x_i$

Sample Variance: $\sigma_k^2 = \frac{1}{d * N_k}\sum_{i=1}^{N_k} ||x_i - \mu_k||^2$

Gaussian-Gaussian

Exponential: $\lambda exp(-\lambda x_i)$

Gamma: $\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{\beta\lambda}$

- constant: $\frac{\beta^\alpha}{\Gamma(\alpha)}$

$$\alpha^{t+1} = \alpha^t + N$$

$$\beta^{t+1} = \beta^t + \sum_{i=1}^{N} x_i$$

Exponential-Gamma

Naive Bayes: $\frac{P(\mathbf{x}^*|C_1)P(C_1)}{P(\mathbf{x}^*|C_2)P(C_2)} \underset{C_2}{\overset{C_1}{\gtrless}} 1$

Discriminant Function: $ln[\frac{P(\mathbf{x}^*|C_1)P(C_1)}{P(\mathbf{x}^*|C_2)P(C_2)}] \underset{C_2}{\overset{C_1}{\gtrless}} 0$

$$g(x) = ln(g_1(x)) - ln(g_2(x)) \underset{C_2}{\overset{C_1}{\gtrless}} 0$$

Mixture Models (GMM):

$$L = \sum_{i=1}^{N} ln(\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k))$$

- $\sum \pi_k = 1$

GMM: not sensitive to scaling
Integer encoding: need scaling

GMM- EM: $\sum_{z_i=1}^{K} ln(\prod_{i=i}^{N} \pi_{zi} N(x_i|\mu_{zi}, \Sigma_{zi}))C_{ik}$

- C_{ik} is the membership

$$\mu_k = \frac{\sum_{i=1}^N \mathbf{x}_i C_{ik}}{\sum_{i=1}^N C_{ik}}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^N ||x_i - \mu_k||_2^2 Cik}{d * \sum_{i=1}^N C_{ik}}$$

$$\pi_k = \frac{\sum_{i=1}^N C_{ik}}{N}$$

$$s = \frac{1}{N}\sum_{i=1}^{N} \frac{b_i - a_i}{\max(a_i, b_i)}$$

Silhouette Index:

- a_i is the average distance of the point x_i to all the other points of the cluster in which x_i is assigned to
- b_i is the average distance of the point x_i to all the other points in the other clusters.

Rand Index:

$$r = \frac{a+b}{a+b+c+d}$$

- a is the number of pairs of elements in X that are in the same subset in C and in the same subset in D.
- b is the number of pairs of elements in X that are in different subset in C and in different subset in D.
- c is the number of pairs of elements in X that are in the same subset in C and in different subset in D.
- d is the number of pairs of elements in X that are in different subset in C and in the same subset in D.

KMeans: $J(\Theta, U) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik} d^2(x_i, \theta_k) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik} ||x_i - \theta_k||_2^2$

- Hard Assignment: $u_{ik} \in \{0, 1\} \text{ and } \sum_{k=1}^{K} u_{ik} = 1$
- Lagrange: $\sum_{i=1}^{N} \lambda_i (1 - \sum_{j=1}^{K} u_i j)$

- Mahalanobis Distance(d): $d^2(x_i, \theta_j) = (x_i - \theta_j)^T \Sigma^{-1} (x_i - \theta_j)$

class $\theta_k = \frac{\sum_{x_i \in C_k} x_i}{N_k}$ KNN: Choose closest of highest probability

Unweighted: $P(C_i) = \frac{N_i}{N}$

- N_i are the neighbors for each class

Weighted:

$$\sum_i \frac{1}{d_i}$$

- d_i is a distance formula
- Euclidian: $d_E = \|\mathbf{x}_1 - \mathbf{x}_2\|_2 = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)}$
- Mahalanobis: $d_M = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$
- City-block: $d_{CB} = \sum_{i=1}^{n} |\mathbf{x}_{1i} - \mathbf{x}_{2i}|$
- Cosine: $d_{cos} = 1 - \cos(\angle(x_1, x_2)) = 1 - \frac{\mathbf{x}_1^T \mathbf{x}_2}{\|\mathbf{x}_1\|_2 \|\mathbf{x}_2\|_2}$

Supervised learning: Data collection, Feature extraction, Mapper function, Objective function, Learning algorithm

Overfitting: more data, Occam’s Razor, Cross-validation, Regularization

MAE: Can mitigate the impacts of outliers

MSE: heavily penalize large error

CV: stratified when imbalance proportion