ML Algorithms

Christopher Chan

January 23, 2019

Algorithms

- Linear Regression
- Logistic Regression
- Naive Bayes
- Trees
- SVM

Linear Regression

Assumptions:

1. Observations y_i are uncorrelated
2. Observations y_i have constant variance
3. x_i are fixed

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$$

Cost function

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$X = N * (p+1)$$

Normal equation

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- With

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

Gradient descent

$$\beta_j := \beta_j - \alpha \frac{\partial}{\partial \beta_j} (\beta_0 ... \beta_p)$$

Implement a gradient descent algorithm Gradient descent for β_j

$$\beta_j := \beta_j - \alpha \frac{1}{p} \sum_{i=1}^{p} (\sum x_j \beta_j)$$

Hat (Matrix) - predicts y

$$H = X(X^T X)^{-1} X^T$$

Variance - unbiased

$$\hat{\sigma} = \frac{1}{N-p-1} \sum_{i=1}^{N} (y_i - \hat{y_i})^2$$

Ridge Regression

$$\hat{\beta}^{ridge} = argmin_{\beta}[\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2]$$

Lasso Regression

$$\hat{\beta}^{lasso} = argmin_{\beta}[\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|]$$

Logistic Regresion

In general modeling:

$$p(X) = p(Y = y|X)$$

log-odds

$$log\frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}$$

Estimating regression coefficients with likelihood function:

$$\ell(\beta_0, \beta_1 ... \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_i'=0} (1 - p(x_{i'}))$$

Cost function:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} log(h_\theta(x^{(i)})) + (1-y^{(i)}) log(1 - h_\theta(x^{(i)}))]$$

K Nearest Neighbor

Naive Bayes

$$p(y|x_1, ..., x_n) \propto p(y) \prod_{i}^{n} P(x_i|y)$$

$$y = argmax_y P(y) \prod_{i}^{n} P(x_i|y)$$

Trees

Regression

General cost function

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Cost function at each split

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Pruned tree with α cost function Number of elements in each split

$$N_m = \# x_i \in R_m$$

Average y_i for each split

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

MSE at R_m

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$$

Cost complexity with α

$$\mathcal{C}_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha \left| T \right|$$

Classification

Gini index

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk}$$

Information gain

$$IG(T,a) = H(T) - H(T|a) = -\sum_{k=1}^{K} \hat{p}_{mk} log \hat{p}_{mk} - \sum_{a} p(a) \sum_{i=1}^{J} -Pr(i|a) log_2 Pr(i|a)$$

Bagging

Regression

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Classification

$$\hat{C_{\text{rf}}^{B}}(x) = \text{majority vote } \hat{C}_b(x)_1^B$$

Notes

J = Divides s = cutoff point Divide $X_1, X_2, ...X_p$ into J distinct non-overlapping regions, $R_1, R_2, ..., R_J$

Support Vector Machines

Maximal Marigin classifier

$$\max_{\beta_0,\beta_1,...,\beta_p,M} M$$

subject to $\sum_{j=1}^{p} \beta_j^2 = 1$, $y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \geq M \forall i = 1, ..., n$

Support vector classifier

$$\max_{\beta_0,\beta_1,...,\beta_p,\epsilon_1,...,\epsilon_n,M} M$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1, y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \geq M(1 - \epsilon_i), \epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Alternative formula:

$$\min ||\beta|| \text{subject to} \begin{cases} y_i(x_i^T \beta + \beta_0) \geq 1 - \epsilon_i \forall, \\ \epsilon_i \geq 0, \sum \epsilon_i < \text{constant} \end{cases}$$

- C is a nonnegative tuning parameter
- $\epsilon_1, ..., \epsilon_n$ are slack variables
    - If $\epsilon_i > 0$, then ith observation on wrong side of margin
    - If $\epsilon_i > 1$, then ith observation on wrong side of hyperplane

Who is who:

Symbol	Meaning
N	Number of observations
p	Number of parameters
β	GLM coefficient
$p()$	Probability
J	Distinct, non-overlapping regions in a tree model
s	Cutoff point between regions in tree model
$\hat{C}_b(x)$	Class prediction of bth rf

Notes

x parameterized by θ:

$$x; \theta$$