Formula Sheet — Machine Learning 1

Master Data Science — Summer Semester 2019

Tim Downie — Last edit: June 17, 2019

K-Means Clustering Minimise $W(C_1, \dots C_k) = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x_i} - \boldsymbol{\mu_k}\|^2$,

$\boldsymbol{\mu_k} = (\mu_{k,1}, \dots \mu_{k,p})$, with $\mu_{k,j} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{i,j}$ and
$\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \sum_{j=1}^{p} (x_j - y_j)^2$ is the squared *euclidean distance* between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$

For K-Medians use Manhattan distance $\|\boldsymbol{x} - \boldsymbol{y}\| = \sum_{j=1}^{p} |x_j - y_j|$.

Mean Square Error (MSE)

$$\mathrm{MSE}(f, \boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 = \frac{1}{n} RSS \text{ (Residual sum of squares)}$$

Regression

Linear Models A linear model has the form $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$.
In vector and matrix form: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. $\boldsymbol{X}$ is called the design matrix and $\boldsymbol{\beta}$ is the vector of parameters. The linear model estimates are the least squares estimates $\widehat{\boldsymbol{\beta}}$, which minimises the residual sum of squares.
$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$. Fitted values: $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ Residuals: $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$

An **Anova model** is a linear model where the explanatory variables are discrete or nominal. Example one factorial Anova model, an overall parameter β_0 is fitted and then one parameter is fitted for all but the first level.

$$y_i = \beta_0 + \beta_1 \mathcal{I}(x_i \in A_2) + \cdots + \beta_{K-1} \mathcal{I}(x_i \in A_K) + \epsilon_i$$

$\mathcal{I}(x_i \in A_k)$ is equal to 1 if x_i is in the group A_k and zero if not. A_1 is called the base line group.

Ridge Regression minimises

$$RSS + \lambda \sum_{j=1}^{p} \widehat{\beta}_j^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j x_i \right)^2 + \lambda \sum_{j=1}^{p} \widehat{\beta}_j^2$$

The Lasso minimises

$$RSS + \lambda \sum_{j=1}^{p} |\widehat{\beta}_j| = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} \widehat{\beta}_j x_i \right)^2 + \lambda \sum_{j=1}^{p} |\widehat{\beta}_j|$$

Classification

Binary classification Predict $Y = 0$ if $P(Y = 1|x_1, \ \dots, x_p) < \alpha$ else predict $Y = 1$.

Logistic regression

Logit function: $\text{logit}(s) = \log\left(\frac{s}{1-s}\right)$

Logistic regression linear predictor $\text{logit}(f(x_1, \dots, x_p)) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

$$\Rightarrow f(x_1, \dots, x_p) = P(Y = 1|x_1, \ \dots, x_p) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\}}$$

The **sensitivity** of a classifier is the true positive rate. The **specificity** of a classifier is the true negative rate.

The conditional probability of A given that B occurs is $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Bayes Thm 2nd version $P(A|B) = \dfrac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$

Bayes Classifier Choose the group y_j which maximises P$(Y = y_k|x_1, \dots x_p), \ $ for all $k = 1, \dots K$

Bayes Error Rate measures how often we can expect to make a false classification.
$1 - E(\max_k P((Y = y_k|X))$

Linear discriminant analysis (LDA) The LDA classifier rule is to choose the value of k giving the largest $P(Y = y_k|x)$, the posterior probability our outcome variable is in group y_k given x

$$P(Y = y_k|x) = \frac{f_{X|Y=y_k}(x)P(Y = y_k)}{\sum_{j=1}^{K} f_{X|Y=y_j}(x)P(Y = y_j)}$$

$P(Y = y_k)$ is the prior probability that our outcome variable is in group y_k.
The density $f_{X|Y=y_k}(x)$ in each group is modelled using a multinomial distribution:

$$(X_1, X_2, \dots X_p)|Y = y_k \ \sim N(\boldsymbol{\mu_k}, \Sigma^2)$$

$\boldsymbol{\mu_k}$ is a vector of length p. Σ^2 is the common variance-covariance matrix.

Quadratic discriminant analysis (QDA) As LDA but use a different covariance in each group.

$$(X_1, X_2, \dots X_p)|Y = y_k \ \sim N(\boldsymbol{\mu_k}, \Sigma_k^2)$$

Σ_k^2 is the variance-covariance matrix in group k.

Tree Methods

Regression Trees Minimise $RSS = \sum_{j \in |T|} \sum_{i \in R_j} (y_i - \widehat{y}_j)^2$, where the R_js are terminal nodes in the Tree T and $\widehat{y}_j = \dfrac{1}{|R_j|} \displaystyle\sum_{i \in R_j} y_j$ mean of Y in the region R_j.

Pruning The full tree T_0 is be pruned back to the subtree $T_\alpha \subset T_0$ by minimising

$$PLS(\alpha) = \sum_{j\in|T|}\sum_{i\in R_j}(y_i - \widehat{y}_j)^2 + \alpha|T|,$$

with complexity parameter $\alpha \geqslant 0$, and each R_j the terminal nodes in T.

Classification trees

Classification error rate $E = \sum_{m=1}^{M} 1 - \max_k\{\widehat{p}_{mk}\}$

Gini Index $G = \sum_{m=1}^{M}\sum_{k=1}^{K}\widehat{p}_{mk}(1-\widehat{p}_{mk})$

Entropy or Information index $D = -\sum_{m=1}^{M}\sum_{k=1}^{K}\widehat{p}_{mk}\log(\widehat{p}_{mk})$