List of formulas (STK2100)

(Version May 2019)

1 Loss functions

(*a*) For regression, the quadratic loss function is usually used: $L(y, \hat{y}) = (y - \hat{y})^2$. The optimal predictor based on the input variable $\mathbf{x}$ is then $\widehat{Y} = E[Y|\mathbf{x}]$.

(*b*) For classification, the 0-1 loss is usually used: $L(y, \hat{y}) = I(y = \hat{y})$, where $I(\cdot)$ is the indicator function. The optimal predictor based on the input variable $\mathbf{x}$ is then $\widehat{Y} = \text{argmax}_k \Pr(Y = k|\mathbf{x})$.

2 Multi-variable linear regression

(*a*) Model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \,;\; i = 1, 2, \ldots, n \,;$$

where x_{ij}'s are supposed known and ϵ_i's are independent and $N(0, \sigma^2)$-distributed.

(*b*) In matrix form,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ are n- and $(p+1)$-dimensional vectors, respectively, and $\mathbf{X} = \{x_{ij}\}$ (with $x_{i0} = 1$) is an $n \times (p+1)$-dimensional matrix.

(*c*) The ordinary least squares estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(*d*) Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$. Then the $\hat{\beta}_j$'s are unbiased and normally distributed, with

$$\text{Var}(\hat{\beta}_j) = \sigma^2 c_{jj} \quad \text{og} \quad \text{Cov}(\hat{\beta}_j, \hat{\beta}_l) = \sigma^2 c_{jl}$$

where c_{jl} are the element in position (j, l) of the $(p+1) \times (p+1)$ matrix $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$.

(*e*) Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ik}$, and set $\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. Then $S^2 = \frac{\text{SSE}}{n-(p+1)}$ is an unbiased estimator for σ^2, and $[n - (p+1)]S^2/\sigma^2 \sim \chi^2_{n-(p+1)}$. Moreover S^2 and $\hat{\boldsymbol{\beta}}$ are independent.

(*f*) Let $\text{SE}(\hat{\beta}_j)^2$ be the estimator of the variance for $\hat{\beta}_j$ and replace σ^2 with S^2 in the formula for $\text{Var}(\hat{\beta}_j)$ in point (*b*). Then $(\hat{\beta}_j - \beta_j)/\text{SE}(\hat{\beta}_j) \sim t_{n-(p+1)}$.

(*g*) The null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ can be tested by using the test statistic

$$\mathrm{F} = \frac{(\mathrm{SST} - \mathrm{SSE})/p}{\mathrm{SST}/(n-p-1)}$$

where $\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. Under H_0 F is distributed as a Snedecor's F with p and $n-p-1$ degrees of freedom.

(*h*) The null hypothesis

$$H_0 : \beta_{i_1} = \beta_{i_2} = \cdots = \beta_{i_q} = 0$$

can be tested by using the test statistic

$$\mathrm{F} = \frac{(\mathrm{SSE}_0 - \mathrm{SSE})/q}{\mathrm{SSE}/(n-p-1)} \overset{H_0}{\sim} F_{q,n-p-1}$$

where $\mathrm{SSE}_0 = \sum_{i=1}^{n}(y - \hat{y}_i)^2$, with $\hat{y}_i$ computed *under* H_0, while SSE is computed for the full model.

3 The maximum likelihood approach

Assume that $Y_1, Y_2, \ldots, Y_n$ have density $f(y_1, y_2, \ldots, y_n \,|\, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ is a parameter vector (scalar if $d = 1$). Assume that $f(y_1, y_2, \ldots, y_n \,|\, \boldsymbol{\theta})$ satisfies certain regularity conditions.

(*a*) Given the observed values $Y_i = y_i$, $i = 1, \ldots, n$, the likelihood function is $L(\boldsymbol{\theta}) = f(y_1, y_2, \ldots, y_n \,|\, \boldsymbol{\theta})$ and the log-likelihood function $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

(*b*) The maximum likelihood *estimate* is the value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$ or, equivalently, maximizes $l(\boldsymbol{\theta})$. If the observed y_i's are substituted with the stochastic Y_i'c, we get the maximum likelihood *estimator*.

(*c*) The maximum likelihood estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ is a solution of the likelihood equation $s_j(\boldsymbol{\theta}) = 0$, $j = 1, \ldots, d$, where $s_j(\boldsymbol{\theta}) = (\partial/\partial\theta_j) l(\boldsymbol{\theta})$ is the score function. The vector of the score functions is $\mathbf{s}(\boldsymbol{\theta}) = (s_1(\boldsymbol{\theta}), \ldots, s_d(\boldsymbol{\theta}))^T$.

(*d*) The observed information matrix $\bar{\mathbf{J}}(\boldsymbol{\theta})$ is a $d \times d$ matrix with the element (i, j) given by $\bar{J}_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta})$.

The expected information matrix (or Fisher's information matrix) $\bar{\mathbf{I}}(\boldsymbol{\theta})$ is a $d \times d$ matrix with the element (i, j) given by $\bar{I}_{ij}(\boldsymbol{\theta}) = \mathrm{E}[\bar{J}_{ij}(\boldsymbol{\theta})]$.

For independent and identically distributed observations, $\bar{\mathbf{I}}(\boldsymbol{\theta}) = n\mathbf{I}(\boldsymbol{\theta})$, where $\mathbf{I}(\boldsymbol{\theta})$ is the expected information of one observation.

(*e*) When the likelihood equations (at point (*c*)) do not have an explicit solution, the maximum likelihood estimation can be found by using the Newton-Raphson method:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\mathbf{J}}^{-1}(\boldsymbol{\theta}^{(s)})\mathbf{s}(\boldsymbol{\theta}^{(s)}),$$

by using the Fisher scoring algorithm:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\mathbf{I}}^{-1}(\boldsymbol{\theta}^{(s)})\mathbf{s}(\boldsymbol{\theta}^{(s)}),$$

or an appropriate modification of it.

(*f*) For large number of data, $\hat{\theta}_i$ is normally distributed with mean θ_i and variance equal to the i-the diagonal element of $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$. The covariance between $\hat{\theta}_i$ and $\hat{\theta}_j$ is equal to the element (i, j) in $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$. We can estimate the variance/covariance by plugging in $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in $\bar{\mathbf{I}}^{-1}(\boldsymbol{\theta})$ or in $\bar{\mathbf{J}}^{-1}(\boldsymbol{\theta})$.

4 Model selection criteria

(*a*) Degrees of freedom: for linear model, i.e., those for which $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, the number of degrees of freedom is $\text{trace}(\mathbf{S}) = \sum_i S_{ii}$.

(*b*) AIC is defined as $\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2|\boldsymbol{\theta}|$ where $|\boldsymbol{\theta}|$ is the degrees of freedom of the model.

(*c*) BIC is defined as $\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + \log(n)|\boldsymbol{\theta}|$.

5 Some other methods for regression

(*a*) K-nearest neighbor regression is defined by

$$\hat{f}(\mathbf{x}_0) = \frac{1}{K} \sum_{\mathbf{x}_i \in \mathcal{N}_0} y_i$$

where $\mathcal{N}_0 \subset \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ contains the K closest points to $\mathbf{x}_0$.

(*b*) kernel methods (e.g., local regression) are conceptually similar to K-nearest neighbor regression but the influence of an observation depends on (is weighted by) its distance from the point of interest,

$$w_i = \frac{1}{h} w\left(\frac{x_i - x_0}{h}\right)$$

where x_i is the observation and x_0 the point of interest. Typical kernels $w(z)$ are

- Normal, $\frac{1}{\sqrt{2\pi}} \exp\{(-z^2/2)\}$, support $\mathcal{R}$;
- Rectangular, $1/2$, support $(-1, 1)$;

- Epanechnikov, $\frac{3}{4}(1-z^2)$, support $(-1,1)$;
- Biquadratic, $\frac{15}{16}(1-z^2)^2$, support $(-1,1)$;
- Tricubic, $\frac{70}{81}(1-|z|^3)^3$, support $(-1,1)$;

(c) Ridge regression: minimize w.r.t. $\boldsymbol{\beta}$

$$h(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

(d) Lasso regression: minimize w.r.t. $\boldsymbol{\beta}$

$$h(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

(e) Cubic spline: Piecewise polynomials with bases

$$\begin{aligned} b_0(x) =& 1, \quad b_1(x) = x, \quad b_2(x) = x^2, \quad b_3(x) = x^3, \\ b_{3+k}(x) =& (x-c_k)_+^3, \quad k = 1, ..., K \end{aligned}$$

(f) Tree-based methods: $f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m)$ where $\mathcal{R}^p = R_1 \cup R_2 \cup \cdots \cup R_M$ and the regions are defined through sequential splits based on one variable at time.

(g) Bagging and random forest:

$$\hat{f}_{\text{avg}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(\mathbf{x})$$

where $\hat{f}^1(\mathbf{x}), \hat{f}^2(\mathbf{x}), ..., \hat{f}^B(\mathbf{x})$ are B different predictors based on ordinary bootstrapping (bagging) or where the or where only a subset of the explanatory variables are considered in each tree (random forrest).

(h) Neural networks with a latent layer: $f(\boldsymbol{x}) = \beta_0 + \sum_{m=1}^{M}\beta_k\sigma(\alpha_m^T\boldsymbol{x})$.

6 Classification

(a) K-nearest neighbor classification is defined by

$$\Pr(Y = j|\mathbf{X} = \mathbf{x}_0) = \frac{1}{K}\sum_{\mathbf{x}_i \in \mathcal{N}_0} I(y_i = j)$$

where $\mathcal{N}_0 \subset \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ contains the K closest points to $\mathbf{x}_0$.

(*b*) Logistic regression: $Y \in \{0, 1\}$ and

$$\Pr(Y = 1|\mathbf{x}) = \frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T\boldsymbol{\beta}}} = 1 - \Pr(Y = 0|\mathbf{x}).$$

(*c*) Use of Bayes theorem for classification

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

(*i*) LDA: $f_k(\mathbf{x}) = p(x|y = k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$.

(*ii*) QDA: $f_k(\mathbf{x}) = p(x|y = k) = N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

7 Dimensionality reduction

(*a*) Principal components: 1^{st} principal component is defined as $z_1 = \boldsymbol{\phi}_1^T\mathbf{x}$, where $\boldsymbol{\phi}_1$ is chosen such that $\text{var}(z_1)$ is as large as possible.

(*b*) Partial least squares: also uses the response variable to define the transformed variables.

8 Hierarchical clustering

(*a*) Decomposition of the *total dissimilarity*,

$$\sum_{ij} d(i, i') = \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')=k} d(i, i') + \sum_{k=1}^{K} \sum_{G(i)=k} \sum_{G(i')\neq k} d(i, i')$$

where $G(i)$ indicates the cluster that the i-th observation belongs to, K is the total number of clusters and $d(i, i') = \sum_i d_j(x_{ij}, x_{i'j}))$ is the dissimilarity between observations i and i'. Here x_{ij} denotes the j-th component of the observation i.

(*b*) Dissimilarity measures between groups:

- single link, $\min_{i\in G, i'\in G'} d(i, i')$
- complete link, $\max_{i\in G, i'\in G'} d(i, i')$
- average link, $\frac{1}{n_G n_{G'}} \sum_{i\in G} \sum_{i'\in G'} d(i, i')$.