EGZAMIN ZE STATYSTYKI II – 16.06.2005

Zadanie 1.

Udowodnij, że macierz korelacji wektora losowego $X = (X_1, X_2, \ldots, X_p)$ ma nieujemne wartości własne.

Rozwiązanie:
Niech

$$X = (X_1, X_2, \ldots, X_p)^T,\ \tilde{X} = \left(\frac{X_1}{\sigma_1}, \frac{X_2}{\sigma_2}, \ldots, \frac{X_p}{\sigma_p}\right)^T,$$

gdzie $\sigma_i = \sqrt{VarX_i}$. Wtedy

$$\begin{aligned} cov(\tilde{X}) &= \mathbb{E}(\tilde{X} - \mathbb{E}\tilde{X})(\tilde{X} - \mathbb{E}\tilde{X})^T = (\mathbb{E}(\tilde{X}_i - \mathbb{E}\tilde{X}_i)(\tilde{X}_j - \mathbb{E}\tilde{X}_j))_{i,j} \\ &= \left(\frac{\mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)}{\sigma_i \sigma_j}\right)_{i,j} = cor(X). \end{aligned}$$

Ale $cov(\tilde{X}) \geq 0$, bo $0 \leq Var\ t^T\tilde{X} = \mathbb{E}(t^T\tilde{X} - t^T\mathbb{E}\tilde{X})(\tilde{X}^T t - \mathbb{E}\tilde{X}^T t) = t^T cov\tilde{X} t$ dla każdego t. Wartości własne $cor(X)$ są nieujemne ponieważ są równe wariancjom współrzędnych $V^T\tilde{X}$, gdzie V^T jest przekształceniem ortogonalizującym $cov(\tilde{X})$.

Zadanie 2.

Podaj algorytm k-średnich.

Rozwiązanie:
Niech $C : \{1, \ldots, N\} \to \{1, \ldots, K\}$, $K < N$ będzie podziałem N - elementowej populacji na K - klas. Algorytm k - means zaczynając od początkowego C_0, poprawia go w sensie grupowania podobnych obserwacji:

`repeat`
`for (k = 1 .. K)`
$\bar{x}_k = m_k = \arg\min_x \sum_{i:c(i)=k} d(x, x_i)$ // znajdujemy centrum każdej klasy
`for (i = 1 .. N)`
$c(i) = \arg\min_k d(x_i, m_k)$ // znajdujemy nowy, lepszy podział
`until kryterium-stopu`

gdzie $d(x_i, x_j)$ - odległość między obserwacjami i oraz j.

Zadanie 3.

Podaj wzór na przekształcenie ortogonalizujące macierz $X = (x_{ij})_{n \times p}$.

Rozwiązanie:
Niech $X = (x_{ij})_{n \times p}$ - macierz danych, $\bar{x}_{\cdot j}$ - średnia j- tej cechy. Wtedy $\tilde{X} = (x_{ij} - \bar{x}_{\cdot j})_{n \times p}$ - macierz danych scentrowanych. $\frac{1}{n}\tilde{X}^T\tilde{X} =: \Sigma$ - macierz kowariancji danych X. $V = [v_1|\ldots|v_p]$ - macierz wektorów własnych Σ. Wówczas XV jest macierzą obserwacji wektora, którego współrzędne są nieskorelowane.

Zadanie 4.

Wyprowadź wzór na bayesowską regułę klasyfikacyjną w wielomianowym modelu normalnym $(\mathcal{N}(\mu_1, \Sigma_1), \ldots, \mathcal{N}(\mu_k, \Sigma_k))$ z rozkładem a priori $(q_1, \ldots, q_k)$.

Rozwiązanie:
$(q_1, \ldots, q_k)$ - rozkład a priori parametru określającego populację,

$f_i(x) = \frac{1}{\sqrt{det\Sigma_i (2\pi)^p}} exp(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i))$- rozkład i- tej populacji $\mathcal{N}(\mu_i, \Sigma_i)$ na $\mathbf{R}^p$.

Bayesowska reguła klasyfikacyjna klasyfikuje daną obserwację x do tej populacji, która maksymalizuje prawdopodobienstwo a posteriori, tj. $f(i|x) \propto q_i f_i(x)$. Gdy x - ustalone

$$\max_i \ q_i f_i(x) \Leftrightarrow \max_i ln q_i \ - \frac{1}{2} log(det(\Sigma_i)) - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i).$$

Powyższą, kwadratową względem x regułę można uprościć do liniowej, gdy $\Sigma_i = \Sigma \ \forall i$.

Zadanie 5.

Niech Y - zmienna losowa, $X = (X_1, X_2, \ldots, X_p)$ - wetror losowy oraz $covX$ odwracalna. Udowodnij, że $\mathbb{E}(Y - \alpha - \beta^T X)^2$ przyjmuje minimum dla $\beta^* = C^{-1}\sigma_{YX}, \ \alpha^* = \mathbb{E}Y - \beta^{*T}\mathbb{E}X$.

Rozwiązanie:
Chcemy:

$$f(x,\beta) = \mathbb{E}(Y - \alpha - \beta^T X)^2 \to \min_{\alpha,\beta}.$$

Różniczkując:

$$\frac{\mathrm{d}f}{\mathrm{d}\alpha} = -2\mathbb{E}(Y - \alpha - \beta^T X) = 0 \Rightarrow \alpha^* = \mathbb{E}Y - (\beta^*)^T \mathbb{E}X.$$

Mamy już α^*, teraz znajdźmy β^*:

$$\begin{aligned} \nabla_\beta f = 2\mathbb{E}(Y - \alpha - \beta^T X)X^T &= 0 \\ \mathbb{E}(YX^T) - \alpha^* \mathbb{E}X^T - \mathbb{E}\beta^{*T} XX^T &= 0 \\ \mathbb{E}(YX^T) - \mathbb{E}Y\mathbb{E}X^T + (\beta^*)^T \mathbb{E}X\mathbb{E}X^T - \beta^{*T}\mathbb{E}XX^T &= 0. \end{aligned}$$

Stąd

$$\begin{aligned} \sigma_{YX}^T - (\beta^*)^T C &= 0 \\ \beta^{*T} &= \sigma_{YX} C^{-1} \\ \beta^* &= C^{-1}\sigma_{YX}. \end{aligned}$$

Zadanie 6.

Udowodnić, że w rodzinie rozkładów na skończonym $\mathfrak{X}$ rozkład jednostajny ma najwiekszą entropię.

Rozwiązanie:

$$\mathfrak{X} = \{1, \ldots, k\}, \ \ p_i, q_i\text{- gęstości na } \mathfrak{X},$$

Odległość Kullbacka - Leiblera nieujemna z nierówności Jensena:

$$H(p||q) = -\sum_{i=1}^{k} p_i \ log\frac{q_i}{p_i} \geq 0.$$

Entropia p

$$H(p) = -\sum p_i \ log \ p_i \ \leq -\sum p_i \ log \ q_i. \tag{1}$$

Niech $q_i = \frac{1}{k}$ - gęstość rozkładu jednostajnego. Wówczas $\forall p$:

$$-\sum p_i \ log \ q_i \ = -\sum p_i \ log \ \frac{1}{k} = log \ k = H(q),$$

zatem z (1):

$$H(p) \leq H(q),$$

dla każdego p.