模式识别
作业二

181220076, 周韧哲, 本科，人工智能学院，人工智能学院选课

2021 年 6 月 26 日

Problem 1

(a) $\|x_j-\mu_i\|^2$ 代表 x_j 到第 i 组的代表 u_i 的距离, 当 $\gamma_{ij}=0$ 时, 这一项为 0, 说明 $\sum_{i=1}^K\gamma_{ij}\|x_j-\mu_i\|^2$ 就是被分到了第 k 组的 x_j 到与其类代表的距离。从而，$\sum_{j=1}^M\sum_{i=1}^K\gamma_{ij}\|x_j-\mu_i\|^2$ 就代表所有的样本到其对应类代表的距离和。对其最小化也就是让相同组的样本到其类代表的距离和最小，即属于相同组的样本彼此相似，距离最小。所以该式形式化了 K 均值目标。

(b) 设第 k 次迭代后为 γ_{ij}^k,μ_i^k。

i. 固定 u_i 后，$\gamma_{ij}^{k+1}=\arg\min_{\gamma_{ij}^k}\sum_{i=1}^K\sum_{j=1}^M\gamma_{ij}^k\|x_j-\mu_i^k\|^2$。固定 μ_i 后，上式等价于对于每个 x_j，$\gamma_{ij}^{k+1}=\arg\min_{\gamma_{ij}^k}\sum_{i=1}^K\gamma_{ij}^k\|x_j-\mu_i^k\|^2$，即选择距离最小的 μ_i^k 对应类别作为 x_j 的类别。所以

$$\gamma_{ij}^{k+1}=\begin{cases}1, & \|x_j-\mu_j^k\|^2\le\|x_j-\mu_{j'}^k\|^2, j'=1,2,\cdots,K\\ 0, & \text{otherwise}\end{cases}$$

ii. 固定 γ_{ij} 后, $\mu_i^{k+1}=\arg\min_{\mu_i^k}\sum_{i=1}^K\sum_{j=1}^M\gamma_{ij}^{k+1}\|x_j-\mu_i^k\|^2$。$\sum_{j=1}^M\gamma_{ij}^{k+1}\|x_j-\mu_i^k\|^2$ 就是第 i 类样本到 μ_i^k 的距离和, 因而固定 γ_{ij} 后可分别优化每个 μ_i : $\mu_i^{k+1}=\arg\min_{\mu_i^k}\sum_{j=1}^M\gamma_{ij}^{k+1}\|x_j-\mu_i^k\|^2$，由于 $f(x)=\|x\|^2$ 为凸函数，故可求导等于 0 得到最小值

$$u_i^{k+1}=\frac{\sum_{j=1}^M\gamma_{ij}^{k+1}x_j}{\sum_{j=1}^M\gamma_{ij}^{k+1}}$$

(c) 记目标函数为 $J(\gamma,\mu)$，首先证明在 Floyd 算法中 $J(\gamma,\mu)$ 递减。在 i 步，固定 μ 后，由于对于每个样本均归类到离其最近的样本中心，即任意 x_j 都有 $\|x_j-\mu_j^k\|^2\le\|x_j-\mu_{j'}^k\|^2, j'=1,2,\cdots,K$，从而 $J'_{k+1}(\gamma,\mu)\le J_k(\gamma,\mu)$。在 ii 步，固定 γ 后，对于每个 μ_i，都取了 $\sum_{j=1}^M\gamma_{ij}^{k+1}\|x_j-\mu_i^k\|^2$ 的最小值，从而，$J_{k+1}(\gamma,\mu)\le J'_{k+1}(\gamma,\mu)$。又因为 $J(\gamma,\mu)\ge 0$ 显然成立，从而目标函数单调递减且有界，必定收敛。

Problem 2

(a) $\min_\beta\sum_{i=1}^n(y_i-x_i^T\beta)^2$

(b) $\min_\beta (y - X\beta)^T(y - X\beta)$

(c) 将优化项展开，得到

$$\min_\beta y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta$$

由于中间两项为标量，所以可写为 $\min_\beta y^Ty - 2\beta^TX^Ty + \beta^TX^TX\beta$，其为二次规划问题，可直接求梯度等于 0, 得到 $X^TX\beta = X^Ty$，假设 X^TX 可逆，所以 $\beta^* = (X^TX)^{-1}X^Ty$。

(d) 不可逆，当 $d > n$ 时 X^TX 为奇异矩阵。

(e) 该正则项会使得 β 是有偏估计。

(f) $\min_\beta (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$，展开得到

$$\min_\beta y^Ty - y^TX\beta - \beta^TX^Ty + \beta^T(X^TX + \lambda I)\beta$$

加入了 λI 可使得 $X^TX + \lambda I$ 可逆，故解为 $\beta^* = (X^TX + \lambda I)^{-1}X^Ty$

(g) 岭回归得到的解中加入了 λI 使得 $X^TX + \lambda I$ 可逆，是普通线性回归的改良版，计算更可靠。

(h) $\lambda = 0$ 时就是普通线性回归的解，$\lambda = \infty$ 时，正则化项的惩罚最大，故解为 $\beta = 0$。

(i) 不可以。如果联合优化的话，因为 λ 对应项为非负的，如果限制 $\lambda > 0$ 的话，为了最小化目标函数，λ 必然非常小，这样就退化为普通线性回归。如果不限制 λ 的话，为了最小化目标函数，λ 会优化成 $-\infty$，没有实际意义。所以不可以在训练集上联合优化 λ 和 β。

Problem 3

(a) 将 F_β 展开得到

$$F_\beta = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2FN + FP}$$

由于分子分母各项都非负且 $(1+\beta^2)TP \le (1+\beta^2)TP + \beta^2FN + FP$，所以 $0 \le F_\beta \le 1$，$F_\beta = 0$ 取得条件为 $TP = 0$，$F_\beta = 1$ 取得条件为 $FN = FP = 0$。

(b) F_β 对查全率 r 和查准率 p 求偏导得

$$\frac{\partial F_\beta}{\partial r} = \frac{\beta^2p^2}{(\beta^2p + r)^2}, \quad \frac{\partial F_\beta}{\partial p} = \frac{r^2}{(\beta^2p + r)^2}$$

可计算 $\frac{\partial F_\beta}{\partial r} / \frac{\partial F_\beta}{\partial p} = \beta^2\frac{p^2}{r^2}$，容易看出 $\beta > 1$ 时，$\beta^2 > 1$，查全率更重要，$0 \le \beta < 1$ 时，$0 \le \beta^2 < 1$，查准率更重要。

Problem 4

(a) $p(x) = \sum_y p(x, y) = \sum_y p(x|y)p(y) = (\frac{1}{\sqrt{2\pi}})(e^{-2(x+1)^2} + e^{-2(x-1)^2})$

(b) $\mathbb{E}[c_{y,f(x)}] = \sum_{x,y} c_{y,f(x)} p(x,y) = \sum_x \sum_y c_{y,f(x)} p(y|x)p(x) = \sum_x p(x) \sum_y c_{y,f(x)} p(y|x)$，对于每个 x，$\sum_y c_{y,f(x)} p(y|x)$ 是独立的，所以只需最小化 $\sum_y c_{y,f(x)} p(y|x)$。可将其展开：$c_{1,f(x)} p(y = 1|x) + c_{2,f(x)} p(y = 2|x) = \mathbb{I}[f(x) \neq 1] p(y = 1|x) + \mathbb{I}[f(x) \neq 2] p(y = 2|x)$，为了使其最小，只需要比较 $p(y = 1|x)$ 与 $p(y = 2|x)$ 的大小，即 $f(x) = \arg\max_y p(y|x)$，这样能使得上面式子退化为较小的一项。多分类时，此规则仍为最优。代价可写为 $\mathbb{I}[f(x) \neq 1] p(y = 1|x) + \cdots + \mathbb{I}[f(x) \neq C] p(y = C|x)$，同样地 $f(x) = \arg\max_y p(y|x)$ 能使得上式退化为最小的 $C - 1$ 项的和，因而是最优的。

(c) 由贝叶斯公式容易得到 $p(y|x) \propto p(x|y)p(y)$，因此最优分类策略为 $f(x) = \arg\max_y p(y|x) = \arg\max_y p(x|y)p(y)$，$p(x|y)$ 和 $p(y)$ 均可由题中条件写出。

(d) 期望代价可写为 $\mathbb{I}[f(x) \neq 1] 10 p(y = 1|x) + \mathbb{I}[f(x) \neq 2] p(y = 2|x)$，所以最优决策 $f(x)$ 为 $\max(10p(y = 1|x), p(y = 2|x)$ 中对应 y 的取值。

Problem 5

(a) 由于 $U^T U = I, V^T V = I$，所以 $XX^T = U\Sigma V^T V \Sigma^T U^T = U\Sigma\Sigma^T U^T = U\Sigma\Sigma^T U^{-1}$，所以其特征值为 $\Sigma\Sigma^T$ 的对角线元素，设 $r = min(m, n)$，则其特征值为 $\sigma_1^2, \cdots, \sigma_r^2, 0, \cdots, 0$，m-r 个 0，特征向量为 U 的对应列向量。

(b) 类似于 (a)，$X^T X = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T\Sigma V^T = V\Sigma^T\Sigma V^{-1}$，所以其特征值为 $\Sigma^T\Sigma$ 的对角线元素，其特征值为 $\sigma_1^2, \cdots, \sigma_r^2, 0, \cdots, 0$，$n - r$ 个 0，特征向量为 V 的对应列向量。

(c) 有相同的 r 个特征值，其为 $\sigma_1, \ldots, \sigma_r$ 的平方。

(d) XX^T 与 $X^T X$ 的 r 个特征值恰好是 X 的 r 个奇异值的平方。

(e) 我会先计算 XX^T 的特征值，因为这个矩阵维度仅为 2×2，而 $X^T X$ 的特征值就是 XX^T 的特征值再补上 99998 个 0。

Problem 6

(a) 忘记减去平均向量时，第一个特征向量和平均向量之间的 corr 小于 1，较低，减去平均向量后，第一个特征向量和平均向量之间的 corr 较高，等于 1。

(b) scale 变量取值变化时，$e1$ 会变化，而 new_e1 不变，即减去平均向量后，第一个特征向量不随 scale 变量的变化而变化，正确的特征向量是

$$(-0.4158, 0.3331, -0.7253, -0.1940, -0.1857, 0.1073, 0.1481, -0.1735, 0.2108, -0.0994)$$

。