

机器学习导论

综合能力测试

学号, 作者姓名, 邮箱

2017 年 6 月 11 日

1 [40pts] Exponential Families

指数分布族(Exponential Families)是一类在机器学习和统计中非常常见的分布族, 具有良好的性质。在后文不引起歧义的情况下, 简称为指数族。

指数分布族是一组具有如下形式概率密度函数的分布族群:

$$f_X(x|\theta) = h(x) \exp(\eta(\theta) \cdot T(x) - A(\theta)) \quad (1.1)$$

其中, $\eta(\theta)$, $A(\theta)$ 以及函数 $T(\cdot)$, $h(\cdot)$ 都是已知的。

- (1) [10pts] 试证明多项分布(Multinomial distribution)属于指数分布族。
- (2) [10pts] 试证明多元高斯分布(Multivariate Gaussian distribution)属于指数分布族。
- (3) [20pts] 考虑样本集 $\mathcal{D} = \{x_1, \dots, x_n\}$ 是从某个已知的指数族分布中独立同分布地(i.i.d.)采样得到, 即对于 $\forall i \in [1, n]$, 我们有 $f(x_i|\theta) = h(x_i) \exp(\theta^T T(x_i) - A(\theta))$ 。

对参数 θ , 假设其服从如下先验分布:

$$p_\pi(\theta|\chi, \nu) = f(\chi, \nu) \exp(\theta^T \chi - \nu A(\theta)) \quad (1.2)$$

其中, χ 和 ν 是 θ 生成模型的参数。请计算其后验, 并证明后验与先验具有相同的形式。

(Hint: 上述又称为“共轭”(Conjugacy), 在贝叶斯建模中经常用到)

Solution. 此处用于写证明(中英文均可)

2 [40pts] Decision Boundary

考虑二分类问题, 特征空间 $X \in \mathcal{X} = \mathbb{R}^d$, 标记 $Y \in \mathcal{Y} = \{0, 1\}$. 我们对模型做如下生成式假设:

- attribute conditional independence assumption: 对已知类别, 假设所有属性相互独立, 即每个属性特征独立地对分类结果发生影响;
- Bernoulli prior on label: 假设标记满足Bernoulli分布先验, 并记 $\Pr(Y = 1) = \pi$.

(1) [20pts] 假设 $P(X_i|Y)$ 服从指数族分布, 即

$$\Pr(X_i = x_i|Y = y) = h_i(x_i) \exp(\theta_{iy} \cdot T_i(x_i) - A_i(\theta_{iy}))$$

请计算后验概率分布 $\Pr(Y|X)$ 以及分类边界 $\{x \in \mathcal{X} : P(Y = 1|X = x) = P(Y = 0|X = x)\}$. (Hint: 你可以使用sigmoid函数 $\mathcal{S}(x) = 1/(1 + e^{-x})$ 进行化简最终的结果).

(2) [20pts] 假设 $P(\mathbf{X}_i|Y = y)$ 服从高斯分布, 且记均值为 μ_{iy} 以及方差为 σ_i^2 (注意, 这里的方差与标记 Y 是独立的), 请证明分类边界与特征 X 是成线性的。

Solution. 此处用于写解答(中英文均可)

3 [70pts] Theoretical Analysis of k -means Algorithm

给定样本集 $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, k -means 聚类算法希望获得簇划分 $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$, 使得最小化欧式距离

$$J(\gamma, \mu_1, \dots, \mu_k) = \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 \quad (3.1)$$

其中, μ_1, \dots, μ_k 为 k 个簇的中心(means), $\gamma \in \mathbb{R}^{n \times k}$ 为指示矩阵(indicator matrix)定义如下: 若 \mathbf{x}_i 属于第 j 个簇, 则 $\gamma_{ij} = 1$, 否则为 0.

则最经典的 k -means 聚类算法流程如算法 1 中所示(与课本中描述稍有差别, 但实际上是等价的)。

Algorithm 1: k -means Algorithm

1 Initialize μ_1, \dots, μ_k .

2 repeat

3 **Step 1:** Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij} = \begin{cases} 1, & \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

4 **Step 2:** For each $j \in \{1, \dots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j :

$$\mu_j = \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}$$

5 until the objective function J no longer changes;

(1) [10pts] 试证明, 在算法 1 中, **Step 1** 和 **Step 2** 都会使目标函数 J 的值降低.

(2) [10pts] 试证明, 算法 1 会在有限步内停止。

(3) [10pts] 试证明, 目标函数 J 的最小值是关于 k 的非增函数, 其中 k 是聚类簇的数目。

(4) [20pts] 记 $\hat{\mathbf{x}}$ 为 n 个样本的中心点, 定义如下变量,

total deviation	$T(X) = \sum_{i=1}^n \ \mathbf{x}_i - \hat{\mathbf{x}}\ ^2 / n$
intra-cluster deviation	$W_j(X) = \sum_{i=1}^n \gamma_{ij} \ \mathbf{x}_i - \mu_j\ ^2 / \sum_{i=1}^n \gamma_{ij}$
inter-cluster deviation	$B(X) = \sum_{j=1}^k \frac{\sum_{i=1}^n \gamma_{ij}}{n} \ \mu_j - \hat{\mathbf{x}}\ ^2$

试探究以上三个变量之间有什么样的等式关系? 基于此, 请证明, k -means 聚类算法可以认为是在最小化 intra-cluster deviation 的加权平均, 同时近似最大化 inter-cluster deviation.

(5) [20pts] 在公式(3.1)中, 我们使用 ℓ_2 -范数来度量距离(即欧式距离), 下面我们考虑使用 ℓ_1 -范数来度量距离

$$J'(\gamma, \mu_1, \dots, \mu_k) = \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \|\mathbf{x}_i - \mu_j\|_1 \quad (3.2)$$

- [10pts] 请仿效算法1(k -means- ℓ_2 算法), 给出新的算法(命名为 k -means- ℓ_1 算法)以优化公式3.2中的目标函数 J' .
- [10pts] 当样本集中存在少量异常点(outliers)时, 上述的 k -means- ℓ_2 和 k -means- ℓ_1 算法, 我们应该采用哪种算法? 即, 哪个算法具有更好的鲁棒性? 请说明理由。

Solution. 此处用于写解答(中英文均可)

4 [50pts] Kernel, Optimization and Learning

给定样本集 $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, $\mathcal{F} = \{\Phi_1 \dots, \Phi_d\}$ 为非线性映射族。考虑如下的优化问题

$$\min_{\mathbf{w}, \mu \in \Delta_q} \frac{1}{2} \sum_{k=1}^d \frac{1}{\mu_k} \|\mathbf{w}_k\|_2^2 + C \sum_{i=1}^m \max \left\{ 0, 1 - y_i \left(\sum_{k=1}^d \mathbf{w}_k \cdot \Phi_k(\mathbf{x}_i) \right) \right\} \quad (4.1)$$

其中, $\Delta_q = \{\mu | \mu_k \geq 0, k = 1, \dots, d; \|\mu\|_q = 1\}$.

(1) [30pts] 请证明, 下面的问题4.2是优化问题4.1的对偶问题。

$$\begin{aligned} \max_{\alpha} \quad & 2\alpha^T \mathbf{1} - \left\| \begin{array}{c} \alpha^T \mathbf{Y}^T \mathbf{K}_1 \mathbf{Y} \alpha \\ \vdots \\ \alpha^T \mathbf{Y}^T \mathbf{K}_d \mathbf{Y} \alpha \end{array} \right\|_p \\ \text{s.t.} \quad & \mathbf{0} \leq \alpha \leq \mathbf{C} \end{aligned} \quad (4.2)$$

其中, p 和 q 满足共轭关系, 即 $\frac{1}{p} + \frac{1}{q} = 1$. 同时, $\mathbf{Y} = \text{diag}([y_1, \dots, y_m])$, \mathbf{K}_k 是由 Φ_k 定义的核函数(kernel).

(2) [20pts] 考虑在优化问题4.2中, 当 $p = 1$ 时, 试化简该问题。

Solution. 此处用于写解答(中英文均可)