

机器学习笔记

张鑫

1 概论

1.1 统计学习三要素

统计学习三要素：模型、策略、算法。在监督学习过程中，模型就是所要学习的条件概率分布或决策函数。模型的假设空间包含所有可能的条件概率分布或决策函数。统计学习的目标在于从假设空间中选取最优模型。监督学习的两个基本策略：经验风险最小化和结构风险最小化。统计学习的算法为求解最优化问题的算法。

1.2 模型评估与模型选择

1.3 机器学习的可能性

1.3.1 Hoeffding 不等式

2 线性回归

2.1 Cost Function

$$h_{\theta}(x) = \theta^T x$$

Assume:

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

where $\epsilon^{(i)}$ is an error term. Further assume $\epsilon^{(i)}$ is i.i.d. and with a Gaussian distribution of mean 0 and variance σ^2 :

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\epsilon^{(i)})^2}{2\sigma^2}}$$

This implies that

$$p(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}$$

Likelihood function:

$$\begin{aligned} L(\theta) &= L(\theta; X, y) = p(y|X; \theta) \\ &= \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \theta) \\ &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}} \end{aligned}$$

Log likelihood:

$$\begin{aligned} l(\theta) &= \log(L(\theta)) \\ &= \log\left(\prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}\right) \\ &= \sum_{i=1}^m \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}\right) \end{aligned}$$

对数似然函数最大化等价于最小化如下 cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (y^{(i)} - h_{\theta}(x^{(i)}))^2$$

2.2 LMS algorithm

只有一个训练样本 (x, y) 时的导数:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (y - h_{\theta}(x))^2 \\ &= (h_{\theta}(x) - y) \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^n \theta_i x_i - y \right) \\ &= (h_{\theta}(x) - y) x_j \end{aligned}$$

相应的更新规则:

$$\theta_j := \theta_j - \alpha (h_{\theta}(x^{(i)}) - y^i) x_j^{(i)}$$

多个样本的更新规则 (Batch gradient descent):

$$\theta_j := \theta_j - \alpha \sum_{i=0}^n (h_{\theta}(x^{(i)}) - y^i) x_j^{(i)}$$

2.3 Norm Equations

迹技巧: If A and B are square matrices and a is a real number,

$$\text{tr } AB = \text{tr } BA$$

$$\text{tr } A = \text{tr } A^T$$

$$\text{tr}(A + B) = \text{tr } A + \text{tr } B$$

$$\text{tr } aA = a \text{tr } A$$

Matrix derivatives:

$$\nabla_A \text{tr } AB = B^T$$

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T$$

$$\nabla_A \text{tr } ABA^T C = CAB + C^T AB^T$$

证明 2.1 假设 A 是 $m \times n$, B 是 $n \times n$, C 是 $m \times m$, 则

$$\begin{aligned} \text{tr } ABA^T C &= \sum_{i=1}^m (ABA^T C)_{ii} \\ &= \sum_{i=1}^m \sum_{j=1}^m (ABA^T)_{ij} C_{ji} \\ &= \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n (AB)_{ik} A_{kj}^T C_{ji} \\ &= \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n \sum_{p=1}^n A_{ip} B_{pk} A_{kj}^T C_{ji} \\ &= \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n \sum_{p=1}^n A_{ip} B_{pk} A_{jk} C_{ji} \end{aligned}$$

$$\begin{aligned}
\frac{\partial \operatorname{tr} ABA^T C}{\partial A_{mn}} &= \frac{\partial (\sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^n \sum_{p=1}^n A_{ip} B_{pk} A_{jk} C_{ji})}{\partial A_{mn}} \\
&= \sum_{k=1}^n \sum_{j=1}^m B_{nk} A_{jk} C_{jm} + \sum_{p=1}^n \sum_{i=1}^m A_{ip} B_{pn} C_{mi} \\
&= \sum_{j=1}^m \sum_{k=1}^n B_{nk} A_{kj}^T C_{jm} + \sum_{i=1}^m \sum_{p=1}^n A_{ip} B_{pn} C_{mi} \\
&= \sum_{j=1}^m (BA^T)_{nj} C_{jm} + \sum_{i=1}^m (AB)_{in} C_{mi} \\
&= \sum_{j=1}^m C_{mj}^T (AB^T)_{jn} + (ABC)_{mn} \\
&= (C^T AB^T)_{mn} + (ABC)_{mn}
\end{aligned}$$

则

$$\nabla_A \operatorname{tr} ABA^T C = C^T AB^T + ABC$$

$$\nabla_A |A| = |A|(A^{-1})^T$$

证明 2.2

$$\begin{aligned}
\frac{\partial |A|}{\partial A_{mn}} &= \frac{\partial \sum_{i=1}^n A_{in} C_{in}}{\partial A_{mn}} \\
&= C_{mn}
\end{aligned}$$

其中, C 为 A 的余子矩阵. 又因为

$$AC^T = (\det A)I,$$

$$C = (\det A)(A^{-1})^T,$$

则

$$\nabla_A |A| = |A|(A^{-1})^T$$

$$X\theta - y = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ h_\theta(x^{(2)}) - y^{(2)} \\ \vdots \\ h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix}$$

$$\begin{aligned}
J(\theta) &= \frac{1}{2}(X\theta - y)^T(X\theta - y) \\
\nabla_{\theta}J(\theta) &= \frac{1}{2}\nabla_{\theta}(X\theta - y)^T(X\theta - y) \\
&= \frac{1}{2}\nabla_{\theta}(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y) \\
&= \frac{1}{2}((\nabla_{\theta^T}\theta^T(X^T X)\theta I)^T - (\nabla_{\theta^T}\theta^T X^T y)^T - (\nabla_{\theta^T}y^T X\theta)^T) \\
&= \frac{1}{2}((\nabla_{\theta^T}\text{tr}\theta^T(X^T X)\theta I)^T - (\nabla_{\theta^T}\text{tr}\theta^T X^T y)^T - (\nabla_{\theta^T}\text{tr}\theta^T X^T y)^T) \\
&= \frac{1}{2}((\theta^T X^T X + \theta^T X^T X)^T - X^T y - X^T y) \\
&= X^T X\theta - X^T y
\end{aligned}$$

令 $\nabla_{\theta}J(\theta) = 0$, 则可得到 normal equation:

$$\theta = (X^T X)^{-1}X^T y$$

3 Logistic Regression

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

where

$$\begin{aligned}
g(z) &= \frac{1}{1 + e^{-z}} \\
g'(z) &= g(z)(1 - g(z))
\end{aligned}$$

Assume that

$$\begin{aligned}
P(y = 1|x; \theta) &= h_{\theta}(x) \\
P(y = 0|x; \theta) &= 1 - h_{\theta}(x)
\end{aligned}$$

等同于

$$p(y|x; \theta) = (h_{\theta}(x))^y(1 - h_{\theta}(x))^{(1-y)}$$

Assuming that the m training examples were generated independently, then the likelihood of the parameters is

$$L(\theta) = \prod_{i=1}^m (h_{\theta}(x^{(i)}))^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{(1-y^{(i)})}$$

The log likelihood is

$$l(\theta) = \sum_{i=1}^m \log((h_{\theta}(x^{(i)}))^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{(1-y^{(i)})})$$

定义损失函数

$$\begin{aligned} J(\theta) &= -l(\theta) \\ &= \sum_{i=1}^m -\log((h_{\theta}(x^{(i)}))^{y^{(i)}} (1 - h_{\theta}(x^{(i)}))^{(1-y^{(i)})}) \\ &= \sum_{i=1}^m -y^{(i)} \log(h_{\theta}(x^{(i)})) - (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)})) \end{aligned}$$

导数为

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta_j} &= \sum_{i=1}^m -y^{(i)} \frac{h_{\theta}(x^{(i)})(1 - h_{\theta}(x^{(i)}))}{h_{\theta}(x^{(i)})} x_j^{(i)} - (1 - y^{(i)}) \frac{-h_{\theta}(x^{(i)})(1 - h_{\theta}(x^{(i)}))}{1 - h_{\theta}(x^{(i)})} x_j^{(i)} \\ &= \sum_{i=1}^m (-y^{(i)}(1 - h_{\theta}(x^{(i)})) + (1 - y^{(i)})h_{\theta}(x^{(i)})) x_j^{(i)} \\ &= \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \end{aligned}$$

多个样本的更新规则 (和线性回归更新规则相同):

$$\theta_j := \theta_j - \alpha \sum_{i=0}^n (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

4 广义线性模型

The exponential family:

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta))$$

η is the *natural parameter* (or *canonical parameter*); $T(y)$ is the *sufficient statistic*; $a(\eta)$ is the *log partition function*.

5 Generative Learning Algorithm

5.1 Gaussian Discriminant Analysis

6 感知机

6.1 感知机学习策略

感知机是一种线性分类模型，属于判别模型。

假设训练数据集线性可分，输入空间 R^n 中任一点到超平面 S 的距离为：

$$-\frac{1}{\|w\|} y_i (w^T \cdot x_i + b)$$

假设超平面 S 的误分类点所有误分类点到超平面 S 的总距离为：

$$-\frac{1}{\|w\|} \sum_{x_i \in M} y_i (w^T \cdot x_i + b)$$

不考虑 $\frac{1}{\|w\|}$ ，得到感知机学习的损失函数：

$$L(w, b) = - \sum_{x_i \in M} y_i (w^T \cdot x_i + b)$$

6.2 感知机学习算法

6.2.1 算法的收敛性

定理 6.1 设训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 是线性可分的，则

(1) 存在满足条件 $\|\hat{w}_{opt}\| = 1$ 的超平面将 T 完全正确分开；且存在 $\gamma > 0$ ，对所有 $i = 1, 2, \dots, N$

$$y_i (\hat{w}_{opt} \cdot \hat{x}_i) = y_i (w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

(2) 令 $R = \max_{1 \leq i \leq N} \|\hat{x}_i\|$ ，则感知机算法在 T 上的误分类次数 k 满足不等式

$$k \leq \left(\frac{R}{\gamma}\right)^2$$

7 Support Vector Machine

$$h_{w,b}(x) = g(w^T x + b).$$

Here, $g(z) = 1$ if $z > 0$, and $g(z) = -1$ otherwise.

Functional margin with respect to the training example $(x^{(i)}, y^{(i)})$:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

Define the function margin of (w, b) with respect to training set to be the smallest of the functional margins of the individual training examples:

$$\hat{\gamma} = \min_{i=1, \dots, m} \hat{\gamma}^{(i)}$$