机器学习复习

2022年4月25日 星期一

CS 229

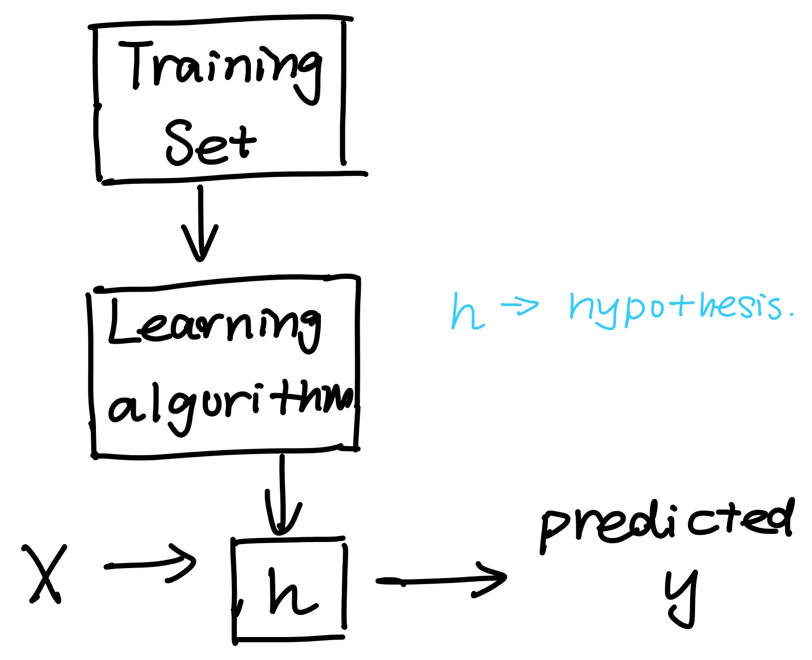
L1. 监督学习，线性回归

① 预测房价模型（提出训练集 $\xrightarrow[h]{\text{算法}}$ 预测模型）

Living Area $(feet)^2$	Price (1000\$s)
2104	400
1600	330
2400	369
⋮	⋮

Put on a plot &

How to predict ?



$h \Rightarrow$ hypothesis.

P1. Line

估计函数：$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

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

错误评价函数：$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

$\min J(\theta)$

How to regulate θ. then get the min of $J(\theta)$?

- 梯度下降法 → gradient descent // 利用 $h(\theta)$，确定 α，迭代找 θ_j

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

  批梯度下降：对全部 Data Set 求误差再更新 θ

  增量梯度下降：每扫描一步对 θ 更新.

- 最小二乘法 → regulate function. // 找到一条直线，使样本点离其距离和取 min.

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

  （X 需要满秩）（shortage: it's slow to reverse a matrix）

- 选用误差函数为平方和的概率解释. // 误差确定 equation

  $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$ ↳ 误差

  $P(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$ // 误差满足正态分布，$\mu = 0$

  最大似然估计为：$L(\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n - \frac{\sum_{i=1}^{m}(y^{(i)} - \theta^T(x^{(i)}))^2}{2\sigma^2}$

  ⇒ 求导结果：$\frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2$

- 带权重的线性回归 // 权重 …

  Basic hypothesis：1. Fit θ to minimize $\sum_i w^{(i)}(y^{(i)} - \theta^T x^{(i)})^2$；2. Output $\theta^T x$

  $w^{(i)}$ predict equation：$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$

  其中 x 为预测特征，离 x 越近，权重越大

- 分类与对数回归

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

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

  // 与线性回归类似，将 $\theta^T x^{(i)}$ 换成了 $h_\theta(x^{(i)})$，$h_\theta(x^{(i)})$ 由 $\theta^T x^{(i)}$ 经 $g(z)$ 映射而形成的

- 牛顿法解最大似然估计 // 解 $f(\theta)=0$ 的方程.

  1.要求解 $f(\theta)=0$ 时，$\theta := \theta - \frac{f(\theta)}{f'(\theta)}$

  2.用于解最大似然估计时，$\theta := \theta - \frac{l'(\theta)}{l''(\theta)}$

  3. θ 为向量时，牛顿法可以使用下式：

  $\theta := \theta - H^{-1}\nabla_\theta l(\theta)$，其中 $H_{ij} = \frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j}$ 是 $n \times n$ 的 Hessian 矩阵 ?

  // 求 Hessian 矩阵比较耗时

  当 x_0 靠近极小值 X 时，牛顿法收敛速度最快，但当 x_0 远离极小值时，牛顿法可能不收敛，甚至不能下降.

  ∵因为 x_{k+1} 不一定为 f 在牛顿方向上的极小点

- 一般线性模型. // 解释为什么用 $g(z) = \frac{1}{1+e^{-z}}$ 作为二元机

  一个概率分布：$P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$

  伯努利分布（两点分布）的概率表示为：

  $P(y;\phi) = \phi^y(1-\phi)^{1-y}$

  $= \exp(y\ln\phi + (1-y)\ln(1-\phi))$

  $= \exp\left(y\ln\frac{\phi}{1-\phi} + \ln(1-\phi)\right)$

  令 $\eta = \ln\frac{\phi}{1-\phi}$ $\phi = \frac{1}{1+e^{\eta}}$

  $(\eta = \theta^T x)$

- Softmax 回归.

  预测值 y 有 k 种可能，$y \in \{1, 2, \cdots, k\}$

  Def：$\phi_i = P(y = i;\phi)$

  则 $\sum_{i=1}^{k}\phi_i = 1$

  $P(y=k;\phi) = 1 - \sum_{i=1}^{k-1}\phi_i$

  取向量 $T(y)$（k−1 dimension，use $T(y)$ to give $P(y=i)$）

  $T(1) = \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}$ $T(2) = \begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}$ $T(3) = \begin{bmatrix}0\\0\\1\\\vdots\\0\end{bmatrix}$ …

  $T(k-1) = \begin{bmatrix}0\\\vdots\\0\\1\end{bmatrix}$ $T(k) = \begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}$

  适用于一般线性模型：$P(y;\phi) = \phi_1^{T(y)_1}\phi_2^{T(y)_2}\cdots\phi_k^{1-\sum_{i=1}^{k-1}T(y)_i}$

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

  其中：$\eta = \begin{bmatrix}\log(\phi_1/\phi_k)\\\log(\phi_2/\phi_k)\\\vdots\\\log(\phi_{k-1}/\phi_k)\end{bmatrix}$ $a(\eta) = -\log(\phi_k)$ $b(y) = 1$

  $\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k}e^{\eta_j}}$

  最大似然估计：$l(\theta) = \sum_{i=1}^{m}\log(P(y^{(i)}|x^{(i)};\theta))$

  $= \sum_{i=1}^{m}\log\prod_{l=1}^{k}\left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}\right)^{1\{y^{(i)}=l\}}$

  采用 GD 或牛顿迭代求解！