Machine Learning

课件 8

Linear Regression 线性回归

1. Regression 回归

数据点（x1，y1），...,(xn, yn)
xi是位置，通常是向量，有时是标量
yi是数值，通常是标量
目标：找到一个能将位置映射到数值上的函数f

Setup

- data points $(x_1, y_1), \ldots, (x_n, y_n)$
- x_i are locations, usually vectors, sometimes scalars
- y_i are values, usually scalars
- goal: find a function f that maps locations onto values

1.1 什么是简单线性回归()

所谓简单，是指只有一个样本特征，即只有一个**自变量**；所谓线性，是指方程是线性的；所谓回归，是指用**方程来模拟变量之间是如何关联的**。

$$y = \beta_0 + \beta_1 x$$ (假设函数)

a.线性回归就是要找一条直线，并且让这条直线尽可能地拟合图中的数据点。

b.统计的世界不是非黑即白的，它有"灰色地带"，但是统计会将理论与实际间的差别表示出来，也就是**"误差"**。

因此，统计世界中的公式会有一个小尾巴 μ ，用来代表误差，即：

$$y = \beta_0 + \beta_1 x + \mu$$

用最小二乘法拟合模型，即

$$w_{LS} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

为什么是最小二乘法?
为什么不是绝对值?
最小二乘法背后的假设是什么?

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

补　**代价函数**

h(x) 为假设函数（下标 θ 可以忽略）；y 为样本集中的输出值；

x 为样本集中的输入值；m 代表样本数据的总量；

上标 i 表示第几个样本数据。

https://zhuanlan.zhihu.com/p/55307907

$$w_{LS} = \arg\min_{w} \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

同理

a. arg min　　就是使后面这个式子达到最小值时的变量的取值

b .arg min F(x,y)　就是指当 F(x,y)取得最小值时，变量 x,y 的取值

c. (可以先把前面 arg min 忽略) == 上面的 1/2m

d. x_i 为样本集中的输入值

e. w 为假设函数

f. y_i 为样本集中的输出值

为什么是最小二乘法---见机器学习个人笔记完整版 v5.51

Notation with many variations
有许多变化的记号

二．

Linear regression: model specification

线性回归：模型规格 向量 x

1. 单一数据点/

Single data point / linear function单一数据点/线性函数

- location x is a vector位置x是一个向量
- function value at x is modelled as $x^T w$ which is linear in x
- *measured* value y is Gaussian distributed around $x^T w$

$$p(y|x, w) = \mathcal{N}(y|x^T w, \sigma^2) \qquad \text{univariate}$$

- σ^2 is the variance of the measurement noise方差
- value y is scalar标量
- parameter w is unknown未知的
- parameter σ^2 is known已知的
- because $x^T w$ is linear in w this is linear regression

a. 为什么 x 是一个向量

https://zhuanlan.zhihu.com/p/68610306

对一个事物，通过描述其不同特征属性，得到的一行或者一列值，这样

就组成了一个特征值形成的向量

b. 多元线性回归（高斯分布--- > 最小二乘法）

https://zhuanlan.zhihu.com/p/378967282

正态分布 $N(\mu,\sigma^2)$ 推到

单一数据 $p(y|x,w)=\mathcal{N}(y|x^Tw,\sigma^2)$ 小 x

多个数据 $p(y|X,w)=\mathcal{N}(y|Xw,\Sigma)$ 大 x

Towards linear regression for nonlinear functions
迈向非线性函数的线性回归

1. Basis function expansion 基准函数扩展 scalar x 标量 x

$$\phi(x)=[1,x,x^2,\ldots,x^d]^T$$

leads to polynomials in x

$$\phi(x)^Tw=\sum_{i=0}^{d}w_ix^i=w_0+w_1x+w_2x^2+\ldots+w_dx^d$$

- for vector $x=[x_1,x_2]$ the polynomial basis function

$$\phi(x)=[1,x_1,x_2,x_1^2,x_2^2,x_1x_2,\ldots,x_1^d,x_2^d]^T$$

leads to polynomials in x_1 and x_2 (or simply in x):

$$\phi(x)^Tw=\sum_{i+j\le d}w_{ij}x_1^ix_2^j$$

$$=w_{00}+w_{10}x_1+w_{01}x_2+w_{20}x_1^2+w_{02}x_2^2+w_{11}x_1x_2+\ldots+w_{d0}x_1^d+w_{0d}x_2^d$$

相当于对于单个 x　　　　对多个变量

Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

Size (feet2)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

a. Single data point / nonlinear function 单一数据点/非线性函数$p(y|x,w)=\mathcal{N}(y|\phi(x)^T w,\sigma^2)$

b. Multiple data points / nonlinear function

$$p(y|X,w)=\mathcal{N}(y|\phi(X)w,\Sigma)$$

Remember:

Linear regression is linear
because it is linear in the parameters.
线性回归是线性的
因为它在参数上是线性的。

1. Maximum likelihood estimation 最大似然估计

$$\theta_{\mathrm{ML}}=\arg\max_{\theta} p(\mathcal{D}|\theta) \qquad p(\mathcal{D}|\theta)=\prod_{i=1}^{n} p(y_i|x_i,\theta)$$

a. 独立同分布（英语：**Independent and identically distributed**，缩写为 iid、.、）是指一组随机变量中每个变量的概率分布都相同，且这些随机变量互相独立

为什么需要满足i.i.d.假设？

机器学习是利用当前获取到的信息（或数据）进行训练学习，用以对未来的数据进行预测、模拟。所以都是建立在历史数据之上，采用模型去拟合未来的数据。因此需要我们使用的历史数据具有**总体的代表性。**

为什么要有总体代表性？我们要从已有的数据（经验）中总结出规律来对未知数据做决策，如果获取训练数据是不具有总体代表性的，就是特例的情况，那规律就会总结得不好或是错误，因为这些规律是由个例推算的，不具有推广的效果。

通过i.i.d.假设，就可以大大减小训练样本中个例的情形。

b. **最大似然估计和最小二乘估计的区别与联系**

https://blog.csdn.net/xidianzhimeng/article/details/20847289

1.最小二乘估计，最合理的参数估计量应该使得模型能最好地拟合样本数据，也就是估计值和观测值之差的平方和最小

2. 最大似然法，最合理的参数估计量应该使得从模型中抽取该 n 组样本观测值的概率最大，也就是概率分布函数或者说是似然函数最大。

c.

Log-likelihood

$$\begin{aligned}\ell(w) &= \log p(\mathcal{D}|w) = \sum_{i=1}^{n} \log p(y_i|x_i, w) \\ &= \sum_{i=1}^{n} \log \mathcal{N}(y_i|x_i^T w, \sigma^2) \\ &= -\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n}(y_i - x_i^T w)^2}_{\text{mean squared error}} - \frac{n}{2}\log(2\pi\sigma^2)\end{aligned}$$

- mean squared error (MSE) is also called *sum of squared error*, ℓ_2 norm of residual errors, etc.

均方误差 MSE

https://blog.csdn.net/Eric2016_Lv/article/details/52819926

Likelihood

$$p(y|X, w) = \mathcal{N}(y|Xw, \Sigma)$$

Closed-form solution for the ML estimator

$$w_{\mathrm{ML}} = (X^T X)^{-1} X^T y$$

d.

- aka *ordinary least squares* (OLS)

1. 又名普通最小二乘法 OLS

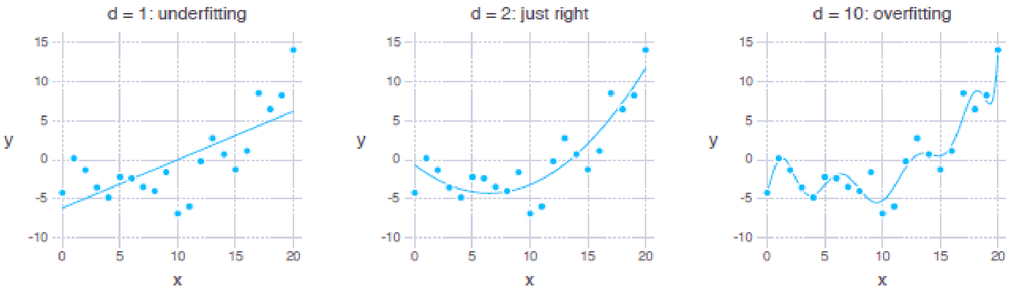
2.正规方程 https://zhuanlan.zhihu.com/p/60719445

e.

$d = 1$ $\quad f(x) = w_0 + w_1 x$

$d = 2$ $\quad f(x) = w_0 + w_1 x + w_2 x^2$

$d = 10$ $\quad f(x) = \sum_{i=0}^{10} w_i x^i$



Notes

- *underfitting* happens if the model is not flexible enough 如果模型不够灵活，就会发生欠拟合。
- *overfitting* happens if the model is too flexible for the amount of data 过度拟合是指模型对所需数量来说过于灵活。

欠拟合 https://zhuanlan.zhihu.com/p/72038532

Ridge regression 岭回归

https://blog.csdn.net/hzw19920329/article/details/77200475

MAP estimation

$$
\begin{aligned}
w_{\text{ridge}} &= \text{argmax}_w\, p(w|X, y) = \text{argmax}_w\, p(y|X, w)p(w|X)/p(y|X) \\
&= \text{argmax}_w\, p(y|X, w)p(w) \\
&= \text{argmax}_w \sum_{i=1}^{n} \log \mathcal{N}(y_i|x_i^T w, \sigma^2) + \sum_{j=1}^{d} \log \mathcal{N}(w_j|0, \tau^2) \\
&= \text{argmin}_w \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T w)^2}_{\text{fit}} + \underbrace{\lambda\|w\|_2^2}_{\text{regularizer}}
\end{aligned}
$$

with $\lambda = \sigma^2/\tau^2$ (just move the σ^2 from the first summand to the second summand and merge with τ^{-2}) and $\|w\|_2^2 = \sum_j w_j^2$.

Solution

$$w_{\text{ridge}} = (\lambda I + X^T X)^{-1} X^T y$$

- d sets model complexity
- λ measures inverse signal-to-noise ratio (next slides)

d设定模型的复杂性

λ 测量反信噪比（下一张幻灯片）。

贝叶斯线性回归

贝叶斯线性回归不仅可以解决极大似然估计中存在的过拟合的问题，而且，它对数据样本的利用率是 100%，仅仅使用训练样本就可以有效而准确的确定模型的复杂度。

Bayesian linear regression (2)

Posterior

$$p(w|X, y) = \mathcal{N}(w|w_n, V_n)$$

$$V_n = (X^T\Sigma^{-1}X + V_0^{-1})^{-1}$$ posterior covariance后验协方差

$$w_n = V_n(V_0^{-1}w_0 + X^T\Sigma^{-1}y)$$ posterior mean后验平均数

Notes

- for $\Sigma = \sigma^2 I$, $V_0 = \tau^2 I$, $w_0 = 0$, the mean of the posterior corresponds to ridge regression后验的平均值对应于岭回归的结果

$$w_n = (\lambda I + X^T X)^{-1} X^T y = w_{\text{ridge}}$$

- however, here we have additionally the posterior covariance

然而，在这里我们还有后验协方差

$$V_n = \sigma^2(\sigma^2/\tau^2 I + X^T X)^{-1} = \sigma^2(\lambda I + X^T X)^{-1}$$

- so λ is the inverse signal-to-noise ratio所以 λ 是反信噪比