Introduction

2017-12-06 14:23

An example

- Hand writing recognition: machine learning is a kind of method of
- Curve Fitting:
  - $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$
  - Root-mean-square error (RMS-Error)
    - $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$
- To prevent over fitting
  - More data
  - By regularization（正则化）
    - $\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$
- Suggest: taking the available data and partitioning it into a training set, used to determine the coefficients w, and a separate validation set, also called a hold-out set, used to optimize the model complexity (either M or λ). However it is a kind of wasting training data
- 这一部分作为导入很贴切，基本说明了是在用什么思路解决什么问题、解决问题中可能遇到的瓶颈以及常用的解决策略

Probability Theory

0.基本概念

- Sum rule of probability (加法法则 边缘概率)
  - $p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$ $= \frac{c_i}{N}.$
- conditional probability (条件概率)
  - $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}.$
- Joint probability (联合概率)
  - $$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\cdot\frac{c_i}{N} \\ &= p(Y = y_j|X = x_i)p(X = x_i) \end{aligned}$$

∵ p(X,Y) = p(X|Y)p(Y)=p(Y|X)p(X)=p(Y,X)

∴ p(Y|X) = p(X|Y)p(Y)/p(X)

- Productive rule (乘法法则)
  - $p(X, Y) = p(Y|X)p(X).$
- Bayes' theorem
  - $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$
- Prior probability (先验概率)：在随机变量的值确定前获得的概率。如从蓝色盒子中抽取的概率是40%
- Posterior probability (后验概率)：在随机变量的值确定后推算出的概率。如已知抽取到的是苹果，推算出之前是从蓝盒子中抽取的概率。

1.概率密度（当随机变量的取值是连续型时）

- The probability that x will lie in an interval (a, b) is then given by
  - $p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x.$
- 原本要求p(y)需要用p_y，在知道随机变量x和y之间的关系后，能将原问题转化成用p_x来求
  - Under a nonlinear change of variable, a probability density transforms differently from a simple function, due to the Jacobian factor. For instance, if we consider a change of variables $x = g(y)$, then a function $f(x)$ becomes $\widetilde{f}(y) = f(g(y))$. Now consider a probability density $p_x(x)$ that corresponds to a density $p_y(y)$ with respect to the new variable y, where the suffices denote the fact that $p_x(x)$ and $p_y(y)$ are different densities. Observations falling in the range $(x, x + \delta x)$ will, for small values of δx, be transformed into the range $(y, y + \delta y)$ where $p_x(x)\delta x \simeq p_y(y)\delta y$, and hence

? 这里的约等于关系是怎么来的

- are different densities. Observations falling in the range $(x, x+\delta x)$ will, for small values of δx, be transformed into the range $(y, y+\delta y)$ where $p_x(x)\delta x \simeq p_y(y)\delta y$, and hence

? 这里的约等于关系是怎么来的

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\,|g'(y)|\,. \end{aligned} \tag{1.27}$$

- 如果随机变量的取值是离散的，那么将p(x)称为概率质量函数probability mass function
- 对于连续型变量，其概率的加法公式和乘法公式为
  - $$\begin{aligned} p(x) &= \int p(x,y)\,\mathrm{d}y \\ p(x,y) &= p(y|x)p(x). \end{aligned}$$

2.Expectations and covariance 期望和协方差

- 在知道一个函数自变量取值的概率分布情况下，可以求该函数的期望
  - $\mathbb{E}[f] = \sum_x p(x)f(x)$
  - $\mathbb{E}[f] = \int p(x)f(x)\,\mathrm{d}x.$
- 方差
  - $\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$
  - $E[(f(x)-E[f(x)])^2] = E[f(x)^2 - 2f(x)E(f(x)) + E[f(x)]^2]$
    $= E[f(x)^2] - 2E[f(x)]^2 + E[f(x)]^2 = E[f(x)^2] - E[f(x)]^2$

3.Bayesian probabilities

- 先验概率 （incorporating with observed data） 后验概率
- $$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1.43}$$

Likelihood function(似然函数)：描述了数据集中对于不同的w观测的情况。

Note：这里的似然函数并非观测在w上的概率分布，所以似然函数累加求和也并不一定等于1。

- 

integrating both sides of (1.43) with respect to w, we can express the denominator in Bayes' theorem in terms of the prior distribution and the likelihood function

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}. \tag{1.45}$$

In both the Bayesian and frequentist paradigms, the likelihood function $p(\mathcal{D}|\mathbf{w})$

- Maximum likelihood(极大似然法): in which w is set to the value that maximizes the likelihood function p(D|w)

4.Gaussian distribution (高斯分布 正态分布)

x̄ 方差 σ 标准差

- $$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.46}$$
- $\beta = 1/\sigma^2$, is called the *precision*
-
- $$\mathbb{E}[x] = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)\,x\,\mathrm{d}x = \mu. \tag{1.49}$$
- $$\mathbb{E}[x^2] = \int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)\,x^2\,\mathrm{d}x = \mu^2 + \sigma^2. \tag{1.50}$$

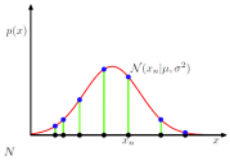
+

$E[(f(x)-E[f(x)])^2]$
$= E[f(x)^2 - 2f(x)E(f(x)) + E[f(x)]^2]$
$= E[f(x)^2] - 2E[f(x)]^2 + E[f(x)]^2$
$= E[f(x)^2] - E[f(x)]^2$

- 如果随机变量x是D维的：

covariance determinant

- $$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.52}$$
- 最大化似然函数，已知x服从正态分布



$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2). \tag{1.53}$$

max

ln Pr

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi). \tag{1.54}$$

直接求偏导，令偏导数为零即可得

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \tag{1.55}$$

ML: max likelihood

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

方差在拟合过程中缩小了

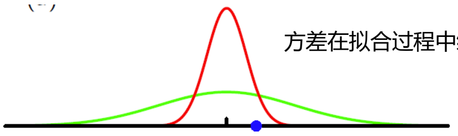
$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \tag{1.57}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \qquad (1.57)$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2 \qquad (1.58)$$

这里并不straigntforword



方差在拟合过程中缩小了

5.Curve fitting re-visited (重新审视曲线拟合) P47