机器学习白板推导系列笔记(1~2.1)

1. 频率派 / 贝叶斯派

机器学习主要解决的是从数据中获取其概率分布的问题. 通过设计机器学习算法从数据中寻找一定的规律, 建立模型来解决实际问题, 因此 ML 问题可以抽象为已知数据集合$\boldsymbol{X}$, $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)^T_{N\times p}$(有$N$个数据每个数据$p$维), 求出相应的参数$\boldsymbol{\theta}$的过程$\boldsymbol{x}\sim p(\boldsymbol{x}|\boldsymbol{\theta})$.

如何求出参数$\boldsymbol{\theta}$, 基本可以分为频率派和贝叶斯派两大学派. 频率派认为参数$\boldsymbol{\theta}$是个确定的常量, 而贝叶斯派认为参数$\boldsymbol{\theta}$也是随机变量, 并且服从某个分布$p(\boldsymbol{\theta})$, 称为先验(prior).

1.1. 频率派

频率派通常使用极大似然估计法(MLE)来求解$\boldsymbol{\theta}$

$$\boldsymbol{\theta}_{MLE}=\arg\max_{\boldsymbol{\theta}}\log P\left(\boldsymbol{X}|\boldsymbol{\theta}\right) \tag{1}$$

对概率取对数是因为$\boldsymbol{X}$中每个样本独立同分布, 概率的连乘取对数后变为连加, 节省运算时间且结果不会因为多个小于 1 的数相乘“消失”.

基于频率派发展出的机器学习方法称为统计机器学习方法, 通常的步骤是建立模型, 定义损失函数(loss function)和最优化损失函数.

1.2. 贝叶斯派

贝叶斯派的方法是基于贝叶斯定理出发的, 在取样结果为$\boldsymbol{X}$时, 其后验概率为

$$P(\boldsymbol{\theta}|\boldsymbol{X})=\frac{P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{X})} \tag{2}$$

其中$P(\boldsymbol{\theta})$为先验概率(prior), $P(\boldsymbol{X}|\boldsymbol{\theta})$为似然概率(likelihood), $P(\boldsymbol{\theta}|\boldsymbol{X})$为后验概率(posterior). 又由全概率公式得

$$P(\boldsymbol{X})=\int_{\boldsymbol{\theta}}P(\boldsymbol{X}\mid\boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\,\boldsymbol{\theta} \tag{3}$$

因此$P(\boldsymbol{\theta}\mid\boldsymbol{X})\propto P(\boldsymbol{X}\mid\boldsymbol{\theta})P(\boldsymbol{\theta})$, 贝叶斯学派一种参数估计的方式是最大后验概率估计(MAP)

$$\boldsymbol{\theta}_{MAP}=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}P(\boldsymbol{\theta}\mid\boldsymbol{X})=\underset{\boldsymbol{\theta}}{\operatorname{argmax}}P(\boldsymbol{X}\mid\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{4}$$

还有一种方式是贝叶斯估计, 它直接求出参数$\boldsymbol{\theta}$的后验概率函数并进行预测估计

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} P(\boldsymbol{X} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{5}$$

假设已经有数据$\boldsymbol{X}$，来了一个新数据$\tilde{\boldsymbol{x}}$，对新数据进行预测

$$P(\tilde{\boldsymbol{x}}|\boldsymbol{X}) = \int_{\boldsymbol{\theta}} P(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} P(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta} \tag{6}$$

注：$P(\tilde{\boldsymbol{x}}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{X}) = P(\tilde{\boldsymbol{x}}|\boldsymbol{\theta}, \boldsymbol{X})P(\boldsymbol{\theta}|\boldsymbol{X})$(因为数据独立同分布)[1]

基于贝叶斯学派发展出的机器学习方法主要是概率图模型.

2. 数学基础

接下来介绍高斯分布，高斯分布在机器学习中有很重要的地位. 假设有如下数据

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^T = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}_{N\times p}$$

其中$\boldsymbol{x}_i \in \mathbb{R}^p$，$\boldsymbol{x}_i \sim (\mu, \Sigma)$(i.i.d)，参数$\boldsymbol{\theta} = (\mu, \Sigma)$.

2.1. 极大似然估计$\boldsymbol{\theta}$

为简单起见，讨论一维高斯分布下的参数估计

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{7}$$

关于$\boldsymbol{\theta}$的似然函数

$$\begin{aligned} \log\ P(\boldsymbol{X} \mid \boldsymbol{\theta}) &= \log \prod_{i=1}^{N} p(x_i \mid \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \log \frac{-1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \sum_{i=1}^{N} \left[\log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sigma} + \frac{(x_i-\mu)^2}{2\sigma^2}\right] \end{aligned}$$

通过极大似然估计求μ_{MLE}

$$\begin{aligned} \mu_{MLE} &= \arg \max_{\mu} \log\ P(\boldsymbol{X} \mid \boldsymbol{\theta}) \\ &= \arg \max_{\mu} \sum_{i=1}^{N} -\frac{(x_i-\mu)^2}{2\sigma^2} \\ &= \arg \min_{\mu} \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

[1] https://zhuanlan.zhihu.com/p/43873322

对似然函数求导$\frac{\partial}{\partial\mu}[\sum_{i=1}^{N}(x_i-\mu)^2]=\sum_{i=1}^{N}2(x_i-\mu)(-1)=0$，解得

$$\mu_{MLE}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{8}$$

通过极大似然估计求σ_{MLE}^2

$$\begin{aligned}\sigma_{MLE}^2&=\arg\max_{\sigma}\log p(\boldsymbol{X}\mid\boldsymbol{\theta})\\&=\arg\max_{\sigma}\sum_{i=1}^{N}\left[-\log\sigma-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\end{aligned}$$

对似然函数求导$\frac{\partial}{\partial\sigma}\sum_{i=1}^{N}\left[-\log\sigma-\frac{(x_i-\mu)^2}{2\sigma^2}\right]=\sum_{i=1}^{N}\left[-\frac{1}{\sigma}+(x_i-\mu)^2\sigma^{-3}\right]=0$，解得

$$\sigma_{MLE}^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 \tag{9}$$

接下来对参数估计结果μ_{MLE},σ_{MLE}^2进行讨论，先上结论：μ_{MLE}是无偏估计，σ_{MLE}^2是有偏估计.

先证明μ_{MLE}是无偏估计，很容易证明

$$\begin{aligned}E[\mu_{MLE}]&=E\left[\frac{1}{N}\sum_{i=1}^{N}x_i\right]\\&=\frac{1}{N}\sum_{i=1}^{N}E[x_i]=\frac{1}{N}\cdot N\cdot\mu=\mu\end{aligned}$$

接着证明σ_{MLE}^2是有偏估计，先对$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$进行适当的变换

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2&=\frac{1}{N}\sum_{i=1}^{N}(x_i^2-2x_i\mu_{MLE}+\mu_{MLE}^2)\\&=\frac{1}{N}\sum_{i=1}^{N}x_i^2-2\frac{1}{N}\sum_{i=1}^{N}x_i\mu_{MLE}+\frac{1}{N}\sum_{i=1}^{N}\mu_{MLE}^2\\&=\frac{1}{N}\sum_{i=1}^{N}x_i^2-2\mu_{MLE}^2+\mu_{MLE}^2\\&=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\mu_{MLE}^2\end{aligned}$$

代入$E[\sigma_{MLE}^2]$得

$$
\begin{aligned}
E[\sigma_{MLE}^2] &= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu_{MLE}^2\right] \\
&= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 - (\mu_{MLE}^2 - \mu^2)\right] \\
&= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] - E[\mu_{MLE}^2 - \mu^2]
\end{aligned}
$$

分别考虑$E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right]$和$E[\mu_{MLE}^2 - \mu^2]$

$$
\begin{aligned}
E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] &= E\left[\frac{1}{N}\sum_{i=1}^{N} \left(x_i^2 - \mu^2\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} E\left[\left(x_i^2 - \mu^2\right)\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} \left(E\left(x_i^2\right) - E^2(x_i)\right) \\
&= \frac{1}{N}\cdot N\cdot\sigma^2 \\
&= \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
E[\mu_{MLE}^2 - \mu^2] &= E[\mu_{MLE}^2] - E(\mu^2) \\
&= E[\mu_{MLE}^2] - \mu^2 \\
&= E[\mu_{MLE}^2] - E^2[\mu_{MLE}] \\
&= \mathrm{Var}\,(\mu_{MLE}) \\
&= \mathrm{Var}\left[\sum_{i=1}^{N}\frac{1}{N}x_i\right] \\
&= \frac{1}{N^2}\mathrm{Var}\left[\sum_{i=1}^{N} x_i\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}\,[x_i](\text{独立同分布}) \\
&= \frac{1}{N^2}N\cdot\sigma^2 \\
&= \frac{1}{N}\sigma^2
\end{aligned}
$$

$E[\sigma_{MLE}^2] = \frac{N-1}{N}\sigma^2$因此对参数σ的参数估计是有偏估计.