Class 6 朴素贝叶斯

2018年11月29日　　16:02

- ## 简介:
  - 　　**朴素贝叶斯（Naive Bayes）算法是机器学习中常见的基本算法之一，它主要被用来做分 类任务。其理论基础是基于贝叶斯定理和条件独立性假设的一种分类方法。**对于给定的训练数据集：

    $$T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$$

    　　首先基于特征条件独立性假设学习联合概率分布 $p(X,Y)$，然后基于此模型，对于任意的输入 x，利用贝叶斯定理求出后验概率最大的$P(Y|X=x)$对应的 y 的取值。

  - 　　基于以上的解释，我们知道：

    （1）该算法的理论核心是贝叶斯定理；

    （2）它是基于条件独立性假设这个强假设基础之上的，这也是它为什么被称为“朴素”的主要原因。

- ## Naive Bayes 算法的数学原理:
  - **贝叶斯定理**

    根据贝叶斯定理，对一个分类问题，给定样本特征 x，样本属于类别 y 的概率是

    $$\boldsymbol{p(y|x)=\frac{p(x,y)\cdot p(y)}{p(x)}}$$

    公式中的 x 是特征向量，假设其维度为 d ，则有

    $$\boldsymbol{p(y|x)=\frac{p(x_1,x_2,\dots,x_d|y)\cdot p(y)}{p(x)}}$$

    朴素贝叶斯法对条件概率分布做了条件独立性的假设，于是有

    $$\boldsymbol{p(x_1,x_2,\cdots,x_d|y=c_k)=\prod_{i=1}^{d}p(x_i|y=c_k)}$$

从而可得朴素贝叶斯分类的基本公式

$$p(y=c_k|x)=\frac{p(y=c_k)\cdot\prod_{i=1}^{d}p(x_i|y=c_k)}{p(x)}$$

- 多项式分布模型：
  - **模型假设**

$$p\left(y=c_j\right)=\pi_j$$

$$p\left(x|c_j\right)=p\left(x_1,x_2,\cdots,x_d|y=c_j\right)=\prod_{i=1}^{d}p\left(x_i|c_j\right)$$
$$=\prod_{i=1}^{V}\theta_{i|j}^{N(t_i,x)}$$

  - **联合概率**

$$p\left(x,y=c_j\right)=p\left(c_j\right)\cdot p\left(x|c_j\right)=\pi_j\prod_{i=1}^{V}\theta_{i|j}^{N(t_i,x)}$$

  - **（联合）似然函数**

$$\begin{aligned}
L(\pi,\theta)&=\log\prod_{k=1}^{N}p(x_k,y_k)\\
&=\log\prod_{k=1}^{N}\sum_{j=1}^{C}I(y_k=c_j)p(y_k=c_j)p(x_k|y_k=c_j)\\
&=\sum_{k=1}^{N}\sum_{j=1}^{C}I(y_k=c_j)\log p(y_k=c_j)p(x_k|y_k=c_j)\\
&=\sum_{k=1}^{N}\sum_{j=1}^{C}I(y_k=c_j)\log\pi_j\prod_{i=1}^{V}\theta_{i|j}^{N(t_i,x_k)}\\
&=\sum_{k=1}^{N}\sum_{j=1}^{C}I(y_k=c_j)\left(\log\pi_j+\sum_{i=1}^{V}N(t_i,x_k)\log\theta_{i|j}\right)
\end{aligned}$$

- **最大似然估计**

$$\max_{\pi,\theta} L(\pi,\theta)$$

$$s.t.\begin{cases}\sum_{j=1}^{C}\pi_j=1\\ \sum_{i=1}^{V}\theta_{i|j}=1, j=1,\dots,C\end{cases}$$

  - 应用拉格朗日乘数：

$$J=L(\pi,\theta)+\alpha(1-\sum_{j=1}^{C}\pi_j)+\sum_{j=1}^{C}\beta_j\,(1-\sum_{i=1}^{V}\theta_{i|j})$$
$$=\sum_{k=1}^{N}\sum_{j=1}^{C}I\left(y_k=c_j\right)[\log\pi_j+\sum_{i=1}^{V}N(t_i,x_k)\log\theta_{i|j}]+\alpha\left(1-\sum_{j=1}^{C}\pi_j\right)+\sum_{j=1}^{C}\beta_j\left(1-\sum_{i=1}^{V}\theta_{i|j}\right)$$

- **闭式MLE解**
  - 梯度

$$\frac{\partial J}{\partial\pi_j}=\sum_{k=1}^{N}I\left(y_k=c_j\right)\frac{1}{\pi_j}-\alpha=0$$

$$\frac{\partial J}{\partial\theta_{i|j}}=\sum_{k=1}^{N}I\left(y_k=c_j\right)\frac{N(t_i,x_k)}{\theta_{i|j}}-\beta_j=0$$

  - MLE解

$$\pi_j=\frac{\sum_{k=1}^{N}I\left(y_k=c_j\right)}{\sum_{k=1}^{N}\sum_{j'=1}^{C}I\left(y_k=c_j\right)}=\frac{N_j}{N}$$

$$\theta_{i|j}=\frac{\sum_{k=1}^{N}I\left(y_k=c_j\right)N(t_i,x_k)}{\sum_{k=1}^{N}I\left(y_k=c_j\right)\sum_{i'=1}^{V}N(t_{i'},x_k)}$$

- **拉普拉斯平滑**
  - 目的：为了防止零概率

$$\pi_j = \frac{\sum_{k=1}^{N} I\left(y_k = c_j\right) + 1}{\sum_{j'=1}^{C} \sum_{k=1}^{N} I\left(y_k = c_j\right) + C}$$

$$\theta_{i|j} = \frac{\sum_{k=1}^{N} I\left(y_k = c_j\right) N(t_i, x_k) + 1}{\sum_{i'=1}^{V} \sum_{k=1}^{N} I\left(y_k = c_j\right) N(t_{i'}, x_k) + V}$$

- 多变量伯努利分布模型:
  - **模型假设**

$$\boldsymbol{p}\left(\boldsymbol{y} = \boldsymbol{c}_j\right) = \boldsymbol{\pi}_j$$

$$\boldsymbol{p}\left(\boldsymbol{x}|\boldsymbol{c}_j\right) = \boldsymbol{p}\left(\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_d|\boldsymbol{c}_j\right) = \prod_{i=1}^{v} \boldsymbol{I}\left(\boldsymbol{t}_i \in \boldsymbol{x}\right) \cdot \boldsymbol{\mu}_{i|j} + \boldsymbol{I}\left(\boldsymbol{t}_i \notin \boldsymbol{x}\right) \cdot \left(\mathbf{1} - \boldsymbol{\mu}_{i|j}\right)$$

  - **联合概率**

$$\boldsymbol{p}\left(\boldsymbol{x}, \boldsymbol{y} = \boldsymbol{c}_j\right) = \boldsymbol{\pi}_j \prod_{i=1}^{v} \boldsymbol{I}\left(\boldsymbol{t}_i \in \boldsymbol{x}\right) \cdot \boldsymbol{\mu}_{i|j} + \boldsymbol{I}\left(\boldsymbol{t}_i \notin \boldsymbol{x}\right) \cdot \left(\mathbf{1} - \boldsymbol{\mu}_{i|j}\right)$$

  - **（联合）似然函数**

$$\begin{aligned}
L(\pi, \mu) &= \log \prod_{k=1}^{N} p(x_k, y_k) \\
&= \sum_{k=1}^{N} \log \sum_{j=1}^{C} I\left(y_k = c_j\right) p(x_k, y_k) \\
&= \sum_{k=1}^{N} \sum_{j=1}^{C} I\left(y_k = c_j\right) \log p(c_j) \prod_{i=1}^{V} I(t_i \epsilon x) p\left(t_i|c_j\right) + I(t_i \notin x)\left(1 - p\left(t_i|c_j\right)\right) \\
&= \sum_{k=1}^{N} \sum_{j=1}^{C} I\left(y_k = c_j\right) \left(\log \pi_j + \sum_{i=1}^{V} I(t_i \epsilon x_k) \log \mu_{i|j} + I(t_i \notin x_k) \log(1 - \mu_{i|j}) \right)
\end{aligned}$$

  - **最大似然估计**

$$\max_{\pi, \mu} L(\pi, \mu)$$

$$s.t.\sum_{j=1}^{C}\pi_j = 1$$

➢ 应用拉格朗日乘数:

$$J = L(\pi,\mu) + \alpha\left(1 - \sum_{j=1}^{C}\pi_j\right)$$
$$= \sum_{k=1}^{N}\sum_{j=1}^{C} I(y_k = c_j)\left(log\pi_j + \sum_{i=1}^{V} I(t_i \epsilon x_k)\, log\mu_{i|j} + I(t_i \notin x)\log(1-\mu_{i|j})\right) + \alpha\left(1 - \sum_{j=1}^{C}\pi_j\right)$$

- **闭式MLE解**
  ➢ 梯度

$$\frac{\partial J}{\partial \pi_j} = \sum_{k=1}^{N} I(y_k = c_j)\frac{1}{\pi_j} - \alpha = 0$$

$$\frac{\partial J}{\partial \mu_{i|j}} = \sum_{k=1}^{N} I(y_k = c_j)\left(\frac{I(t_i \epsilon x_k)}{\mu_{i|j}} - \frac{I(t_i \notin x_k)}{1-\mu_{i|j}}\right) = 0, \forall j = 1, \dots, C.$$

  ➢ MLE解

$$\pi_j = \frac{\sum_{k=1}^{N} I(y_k = c_j)}{\sum_{k=1}^{N}\sum_{j'=1}^{C} I(y_k = c_{j'})} = \frac{N_j}{N}$$

$$\mu_{i|j} = \frac{\sum_{k=1}^{N} I(y_k = c_j)\, I(t_i \epsilon x_k)}{\sum_{k=1}^{N} I(y_k = c_j)}$$

- **拉普拉斯平滑**
  ➢ 目的：为了防止零概率

$$\pi_j = \frac{\sum_{k=1}^{N} I(y_k = c_j) + 1}{\sum_{j'=1}^{C}\sum_{k=1}^{N} I(y_k = c_j) + C}$$

$$\mu_{i|j} = \frac{\sum_{k=1}^{N} I\left(y_k = c_j\right) I(t_i \epsilon x_k) + 1}{\sum_{k=1}^{N} I\left(y_k = c_j\right) + 2}$$