> 整理自同学的笔记。
> 监督学习是指有目标变量或预测目标的机器学习方法，包括分类和回归。

分类中的朴素贝叶斯方法（Naive Bayes Classifier)

以垃圾邮件的分类（Span Detecor）为例。

目标

对于训练过的模型，给定$\mathbf{x}$，给出$P(spam|\mathbf{x})$。

训练数据记作$\{\mathbf{x}_i, y_i\}$， $y_i \in \mathcal{C} = \{spam, not\_spam\}$。

eg: spam email: laptop with the lowest price.

基本假设

1. 属性值x_i条件独立于标签值，即

$$P(x_1, x_2, \cdots, x_{|\mathcal{X}|}|\mathcal{C}) = \prod_i P(x_i|\mathcal{C})$$

   以垃圾邮件分类为例，该问题中的样本$\mathbf{x}_i$ 为表征邮件属性的矢量（比如词向量），表示邮件的整体特征。如果不考虑这一假设，在通常的采样中对$P(\mathbf{x}|c)$的估计往往会导出很小的值（不容易找到两封一样的邮件）。

   而这一假设为我们带来的好处则是摆脱了属性捆绑的桎梏，将单个属性作为统计与概率估计的原子单位，既提高了对数据的利用率也有效地降低了模型需要的参数数目。当然这以真实性为代价。

2. 属性值的分布独立于其出现的位置：

$$P(x_i = w_k|c) = P(x_j = w_k|c), \forall i \neq j$$

   亦即：

$$P(x_i = w_k|c) = P(w_k|c), \forall i$$

   这一条件是我们脱离了对邮件长度与位置的依赖，估计中我们就只需要考虑词频，进一步降低了估计参数的数目和复杂度。

理论依据（贝叶斯定理）

$$\begin{aligned}
\hat{y} &= \arg\max_{c\in\mathcal{C}} P(c|\mathbf{x}) \\
&= \arg\max_{c\in\mathcal{C}} \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})} \\
&= \arg\max_{c\in\mathcal{C}} P(\mathbf{x}|c)P(c) \\
&= \arg\max_{c\in\mathcal{C}} P(c)\prod_i P(x_i|c)\ (assumption\ 1) \\
&= \arg\max_{c\in\mathcal{C}} P(c)\prod_k P(w_k|c)^{t_k}\ (assumption\ 2)
\end{aligned}$$

其中的$P(c)$为先验概率，从采样数据中估计。使先验概率更接近真实分布这一点对采样的多样性提出了一定的要求。

最后的$P(w_k|c)$可以用表示$P(w_k|c) = \frac{n_{ck}}{n_c}$,其中$n_c = \sum_{i:y=c} |x_i|$表示c类出现的次数，$n_{ck}$表示c类中词$w_k$出现的次数。但是注意到如果在采样中只要有$n_{ck} = 0$,那在估计中就一定会有$P(w_k|c) = 0$,这在实际中并不是合理的。为了解决这种问题，有一种方案是Laplace Smoothing:

$$P(w_k|c) = \frac{n_{ck}+1}{n_c+|\mathcal{V}|}$$

朴素贝叶斯分类器训练（Training Naive Bayes Classifier）

Input: trainning samples $\mathcal{D} = \{(\mathbf{x_i}, y_i)\}$
$\mathcal{V} \leftarrow$ the set of distinct words and other tokens in $\mathcal{D}$
for each target value $c \in \mathcal{C}$, do
  $\mathcal{D}_c \leftarrow$ the training samples whose labels are c
  $P(c) \leftarrow \frac{|\mathcal{D}_c|}{|\mathcal{D}|}$
  $T_c \leftarrow$ a single document by concentrating all training samples in $\mathcal{D}_c$
  $n_c \leftarrow |T_c|$
  for $w_k \in \mathcal{V}$ *do*
    $n_{ck} \leftarrow$ the number of times the word w_k occurs in T_c
    $P(w_k|c) = \frac{n_{ck}+1}{n_c+|\mathcal{V}|}$
  endfor
endfor

所谓训练，就是计算$P(w_k|c)$的表罢了。

朴素贝叶斯分类器测试（Testing Naive Bayes Classifier）

Input: A new sample $\mathbf{x}$, 设x_i是$\mathbf{x}$的第 i 个属性, $I = \emptyset$
for $x_1, \cdots, x_i$ do
  if $\exists w_k \in \mathcal{V}$ such that $w_k = x_i$, then
    $I \leftarrow I \cup k$
  end if
end for
predict the label of $\mathbf{x}$ by $\hat{y} = \arg\max_{c\in\mathcal{C}} P(c)\prod_{i\in I} P(w_i|c)$

这个算法虽然简单，但是好用。

算法性能的衡量指标

1. 准确率（Accuracy)

$$\text{Accuracy} = \frac{\#\text{ correctly predicted samples}}{\#\text{ total samples}}$$

   这个指标并不适用于一般情景，它忽略了两种分类错误的不同风险。

2. 查准率（Precision)、召回率（Recall）、F-score：

	T（正确）	F（错误）	总计
P（正例）	TP	FP（第一类错误， 假正例）	正例总数
N（反例）	TN	FN（第二类错误， 假反例）	反例总数
总计	预测正确总数	预测错误总数	样例总数

   则

$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{TP}{TP+FN}$$
$$F_1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

逻辑斯谛回归（Logistic Regression)

目标：

给定集合$\{(\mathbf{x}_i, y_i\}_{i=1}^n$, 其中$y_i \in \{0,1\}$,寻找映射：

$$f: X \to Y, where\ X = (X_1, \cdots, X_d)\ and\ Y \in \{0,1\}$$

基本假设：

1. $Y \sim Bern(P), Y$ 服从伯努利二项分布，$P(Y=1) = p$.
2. $X = (X_1, \cdots, X_d)$中的X_j是连续随机变量。
3. 高斯分布: $P(X_j|Y=0) \sim N(\mu_{j0}, \sigma_j^2), P(X_j|Y=1) \sim N(\mu_{j1}, \sigma_j^2)$
4. X_i, X_j条件独立于Y, $\forall i \neq j$.

理论依据：

综上,

$$\begin{aligned}
P(Y=0|X) &= \frac{P(X|Y=0)P(Y=0)}{P(X|Y=0)P(Y=0) + P(X|Y=1)P(Y=1)} \\
&= \frac{1}{1 + \frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)}} \\
&= \frac{1}{1 + \exp(\ln(\frac{P(X|Y=1)P(Y=1)}{P(X|Y=0)P(Y=0)}))} \\
&= \frac{1}{1 + \exp(\sum_j \ln(\frac{P(X_j|Y=1)}{P(X_j|Y=0)}) + \ln\frac{p}{1-p})} (assumption\ 4)
\end{aligned}$$

而

$$\begin{aligned}
\sum_j \ln(\frac{P(X_j|Y=1)}{P(X_j|Y=0)}) &= \sum_j \ln(\frac{\exp(-\frac{(X_j-\mu_{j1})^2}{2\sigma_j^2})}{\exp(-\frac{(X_j-\mu_{j0})^2}{2\sigma_j^2})}) (assumption\ 3) \\
&= \sum_j \frac{\mu_{j1}-\mu_{j0}}{\sigma_j^2} X_j + \sum_j \frac{\mu_{j0}^2 - \mu_{j1}^2}{2\sigma_j^2}
\end{aligned}$$

将其带回原式，

$$\begin{aligned}
P(Y=0|X) &= \frac{1}{1 + \exp(\sum_j \frac{\mu_{j1}-\mu_{j0}}{\sigma_j^2} X_j + \sum_j \frac{\mu_{j0}^2 - \mu_{j1}^2}{2\sigma_j^2} + \ln\frac{p}{1-p})} \\
&= \frac{1}{1 + \exp(\sum_j w_j X_j + w_0)}
\end{aligned}$$

于是又有

$$P(Y=1|X)=\frac{\exp(\sum_j w_jX_j+w_0)}{1+\exp(\sum_j w_jX_j+w_0)}$$

可见决策平面$\sum_j w_jX_j+w_0=0$是线性的。当找到决策平面时，该分类问题就会迎刃而解。而下一步，我们就需要找出需要的权向量$\mathbf{w}$。

采用最大似然估计法：

$$\begin{aligned}\hat{\mathbf{w}}&=\arg\max_{\mathbf{w}}\prod_i P(y_i|X_i,\mathbf{w})\\&=\arg\max_{\mathbf{w}}\sum_i \ln(P(y_i|X_i,\mathbf{w}))\end{aligned}$$

令$-L(\mathbf{w})=\sum_i(y_i\ln(P(Y=1|X_i,\mathbf{w}))+(1-y_i)\ln(P(Y=0|X_i,\mathbf{w})))$,则问题转化为：

$$\hat{\mathbf{w}}=\arg\min_{\mathbf{w}}L(\mathbf{w})$$

那么似乎可以用梯度下降法来求解该问题。（解的存在性、唯一性（严格凸、强凸））

采用正则化可以保证这两点：

$$\hat{\mathbf{w}}=\arg\max_{\mathbf{w}}L(\mathbf{w})+\frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

对于多分类问题，可以训练多个分类器。其中$Y\in\mathcal{C}=\{c_1,\cdots,c_k\}$，可令

$$P(Y\neq c_k|X)=\frac{1}{1+\exp(\sum_j w_{kj}X_j+w_{k0})}$$

$$P(Y=c_k|X)=\begin{cases}\dfrac{\exp(\sum_j w_{kj}X_j+w_{k0})}{1+\sum_{k=1}^{K-1}\exp(\sum_j w_{kj}X_j+w_{k0})} & k=1,\cdots,K-1\\[2ex]\dfrac{1}{1+\sum_{k=1}^{K-1}\exp(\sum_j w_{kj}X_j+w_{k0})} & k=K-1\end{cases}$$

实际问题：数据的不平衡性

来自不同分类的数据数目不平衡时，回导致训练得出的决策平面有更大的偏移。

解决方案包括：

- undersample（主要）
- oversample