文本分类算法

评价指标

1. 文本自动分类：在给定的分类体系下，根据文本的内容自动地确定文本关联的类别。
   - 基本步骤：训练集实例 -> 预处理 -> 特征选取算法 -> 特征项向量表示 -> 分类算法 -> ...
2. 评价指标：准确率、召回率、F值、宏平均、微平均。

> 对于二分类问题，可将样例按照其真实类别和分类器预测类别划分为：
>
> 真正例（True Positive），真负例（True Negative），假正例（False Positive），假负例（False Negative）
>
> 则**准确率** P、**召回率** R 及F**值**的定义如下：
>
> $$P = \frac{TP}{TP + FP}$$
> $$R = \frac{TP}{TP + FN}$$
> $$F = \frac{2PR}{P + R}$$
>
> 对于多分类问题，可将其分为n个二分类问题，并引入**宏平均**和**微平均**来综合考察。
> 宏平均的准确率$Macro_P$、召回率$Macro_R$及F值$Macro_F$为求各个二分类准确率P_i、召回率R_i及F值F_i的算术平均；
> 而微平均则对每一实例建立全局的混淆矩阵，然后计算其评价指标（$Micro\_P, Micro\_R, Micro\_F$）：
>
> $$Macro\_P = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i} \qquad Micro\_P = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}$$
>
> $$Macro\_R = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} \qquad Micro\_R = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$$
>
> $$Macro\_F = \frac{2Macro\_P \cdot Macro\_R}{Macro\_P + Macro\_R} \qquad Micro\_F = \frac{2Micro\_P \cdot Micro\_R}{Micro\_P + Micro\_R}$$

特征选择

1. 特征选择：
   - 目的：（1）避免过拟合，提高分类准确度；（2）通过降维，大大节省计算时间和空间。

- 方法：（1）文档频率法；（2）信息增益法；（3）互信息法；（4）卡方拟合检验法。

2. 文档频率法：
   - 要点：（1）太频繁的词项没有区分度（为DF设置上限阈值，超过则筛除）；（2）太稀有的词项独立表达的类别信息不强；（3）稀有的词更具有代表性。
   - 优点：易实现、可扩展。
3. 信息增益法：计算某一特征（term）t为整个分类分布所提供的信息量，即计算**不考虑任何特征**的熵与考虑特征t的熵的差值，来表征t出现与否导致的熵的变化。

> 设分类样本$\{c_i\}_{i\in[M]}$服从分布S，对于特征t，其信息增益$\mathrm{Gain}(t)$为：
>
> $$\begin{aligned}\mathrm{Gain}(t) &= \mathrm{Entropy}(S) - \mathrm{Expected\ Entropy}(S_t)\\ &= \left[-\sum_{i=1}^{M}\Pr[c_i]\log\Pr[c_i]\right]\\ &\quad -\left\{\Pr[t]\left[-\sum_{i=1}^{M}\Pr[c_i|t]\log\Pr[c_i|t]\right]+\Pr[\overline{t}]\left[-\sum_{i=1}^{M}\Pr[c_i|\overline{t}]\log\Pr[c_i|\overline{t}]\right]\right\}\end{aligned}$$

   - 不足：只适合用来做“**全局**”的特征选择（即所有类都使用相同的特征集合），无法具体到某个类别上。

4. 互信息法：使用熵的互信息来衡量**特征t和某一类别c的相关程度。**

> 设类别c与特征t，则两者的互信息为：
>
> $$I(t,c)=\log\frac{\Pr[t\wedge c]}{\Pr[t]\Pr[c]}=\log\frac{\Pr[t|c]}{\Pr[t]}$$
>
> 其中$\Pr[t|c],\Pr[t]$可在训练集中通过最大似然估计获得。即假设：
>
> | | c | $\overline{c}$ |
> |---|---|---|
> | t | A | B |
> | $\overline{t}$ | C | D |
>
> 则$I(t,c)\approx\log\frac{A\times(A+B+C+D)}{(A+C)\times(A+B)}$。
>
> 此外，还可以定义特征t对所有类别的平均互信息，以及最大互信息：
>
> $$\begin{aligned}I_{AVG}&=\sum_{i=1}^{m}\Pr[c_i]I(t,c_i)\\ I_{MAX}&=\max_{c_i}\Pr[c_i]I(t,c_i)\end{aligned}$$

   - 特点：（1）$I(t,c)$越大，表示特征t对于类别c的区分能力越强；（2）对同个类别c，相对稀有的词项t会计算得到相对较大的$T(t,c)$。
   - 问题：若一个词的频次不够多，且主要出现在某个类别里，则会出现**较高的互信息**，从而给筛选带来噪音。
     - 解决方法：**先按词频排序，后按互信息大小排序。**

5. 卡方拟合检验法：χ^2统计量用于衡量两种因素的独立性/相关性。**χ^2越大，两者独立性越小，相关性越大。**

> 表格同前。则$\chi^2(t,c)$为：

$$\chi^2(t,c)=\frac{(A+B+C+D)(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

同理可定义：

$$\chi^2_{AVG}(t)=\sum_{i=1}^{m}\Pr[c_i]\chi^2(t,c_i)$$
$$\chi^2_{MAX}(t)=\max_{c_i}\chi^2(t,c_i)$$

分类算法

KNN分类

1. 工作原理：
   - 已知样本数据集，及样本中每个数据对应的类别（标签）；
   - 输入未分类的新数据，将新数据与样本集中数据进行比较，设计算法提取其中最相似的数据（**最近邻**）；
   - 选择前K个最相似的数据，以其中出现次数最多的类别，作为新数据的分类。
2. 加权KNN分类：

   设x为新数据，c为某指定类别，则x归属c的计算方法为：

   $$\operatorname{score}(c|x)=b_c+\sum_{d\in\text{ KNN of }x}\operatorname{sim}(x,d)I(d,c)$$

   其中$\operatorname{sim}(x,d)$为数据x与其邻近数据d的相似度；若d属于类别c，则$I(d,c)=1$，否则为0。
3. 优点：（1）简单有效；（2）重训练代价低；（3）计算时间、空间复杂度与训练规模成线性关系。
4. 不足：（1）不适合在线分类，响应速度慢（KNN是懒惰学习算法）；（2）类别评分非规格化；（3）输出的可解释性弱。

贝叶斯分类

1. 分类思想：利用贝叶斯公式，**通过先验概率和类别的条件概率来估计文档d对类别c_i的后验概率**，以此实现对文档d的分类。

   设c_i为某一分类，E为特征。已知先验概率$\Pr[c_i]$及条件概率$\Pr[E|c_i]$，由贝叶斯公式：

   $$\Pr[c_i|E]=\frac{\Pr[c_i]\Pr[E|c_i]}{\Pr[E]}$$

   若假定样例的特征是独立的，则条件概率进一步可以写为：

   $$\Pr[E|c_i]=\Pr[e_1\wedge e_2\wedge\cdots\wedge e_m|c_i]=\prod_{j=1}^{m}\Pr[e_j|c_i]$$

SVM分类

1. 算法原理：

   假设特征空间上的训练数据集：

> $$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$
> $$x_i\in\mathcal{X}=\mathbb{R}^n,y_i\in\mathcal{Y}=\{1,-1\},i=1,2,\cdots,N$$
>
> 通过间隔最大化或等价地求解相应的**凸二次规划问题**学习得到的分离超平面为：
>
> $$w^*\cdot x+b^*=0$$
>
> 决策函数为$f(x)=\text{sign}(w^*\cdot x+b^*)$。

2. 核心概念：（1）**支持向量（supporting vector）**；（2）**间隔（margin）**。
    - 支持向量：距离超平面最近的点。分割线会根据这些点来确定。
    - 间隔：超平面和距离超平面最近的观测点（即支持向量）之间的距离。有**硬间隔**和**软间隔**之分。
3.