机器学习总结

FU Hanlin

April 3, 2019

Contents

1 特征工程 **1**

2 SVM **3**

Chapter 1

特征工程

1. 特征缩放

    - 线性函数归一化。它对原始数据进行线性变换，使结果映射到 $[0,1]$ 的范围。
      $X_{norm}=\frac{X-X_{min}}{X_{max}-X_{min}}$
      其中 X 为原始数据，X_{max},X_{min} 分别为数据最大值和最小值。
    - 零均值归一化. 它会将原始数据映射到均值为 0，标准差为 1 的分布上。

    - 实际应用中，通过梯度下降法求解的模型通常是需要归一化的，更容易通过梯度下降找到最优解。但对于决策树模型并不适用。

2. Word2Vec

    - CBOW 的目标是根据上下文出现的词语来预测当前词的生成概率，Skip-gram 是根据当前词来预测上下文中各词的生成概率。
    - 神经网络部分：训练权重，使得语料库中所有单词的整体生成概率最大化。
    - 上下文-单词矩阵

3. 避免过拟合的方法

    - 基于模型的方法：
      简化模型（将非线性转化为线性），添加约束项以缩小假设空间（L1/L2 正则），集成学习，Dropout 超参数

- 基于数据的方法：
  1. 一定程度内的随机旋转，平移，缩放，裁剪，填充，左右翻转等。
  2. 对图像中的像素添加噪声扰动。
  3. 颜色变换。
  4. 改变图像的亮度，清晰度，对比度，锐度等。

Chapter 2

SVM

1. 原理
   给定训练样本集，分类学习最基本的想法就是基于训练集 D 再样本空间中找打一个划分超平面，将不同类别的样本分开。
   在样本空间中，划分超平面可通过如下线性方程来描述:

   $$\omega^{\mathrm{T}}x + b = 0$$

   其中 ω 为法向量，决定了超平面的方向；b 为位移项，决定了超平面与原点之间的距离。
   样本中任意点 x 到超平面 (w, b) 的距离可写为

   $$r = \frac{\omega^{\mathrm{T}}x+b}{||\omega||}$$

   假设超平面能将训练样本正确分类，

   $$\omega^{\mathrm{T}}x + b \geq +1, y_i = +1 \ \omega^{\mathrm{T}}x + b \leq -1, y_i = -1$$

   距离超平面最近的这几个训练样本点使等号成立，它们被称为支持向量，两个支持向量到超平面的距离为

   $$\gamma = \frac{2}{||\omega||}$$

   它被称为间隔。欲找到具有最大间隔的划分平面也就是

$$max_{\omega,b}\frac{2}{||\omega||}$$
$$s.t. y_i(\omega^{\mathrm{T}}x+b)\geq +1$$

最大化 $||\omega||^{-1}$，这等价于最小化 $||\omega||^2$，于是可重写为 $min_{\omega,b}\frac{1}{2}||w||^2$

2. 对偶问题
原问题本身是一个凸二次规划问题（目标函数是二次的，约束条件是线性的），能直接用现成的优化计算报求解，但我们有更高效的办法。对式使用拉格朗日乘子法可得到其对偶问题。

$$L(\omega,b,\alpha)=\frac{1}{2}||w||^2+\sum_{m}^{i=1}\alpha_i(1-y_i(\omega^{\mathrm{T}}x+b))$$

令 $L(\omega,b,\alpha)$ 对 ω 和 b 的偏导为零可得

$$\omega=\sum_{i=1}^{m}\alpha_i y_i x_i,$$
$$0=\sum_{i=1}^{m}\alpha_i y_i$$

代入 $L(\omega,b,\alpha)$ 中，

$$L(\omega,b,\alpha)=\frac{1}{2}\omega\omega^{\mathrm{T}}+\sum_{m}^{i=1}\alpha_i-b\sum_{i=1}^{m}\alpha_i y_i-\omega^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i x_i y_i$$

$=\frac{1}{2}\omega\omega^{\mathrm{T}}+\sum_{i=1}^{m}\alpha_i-b\cdot 0-\omega^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i x_i y_i$
$=\frac{1}{2}\omega^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i x_i y_i+\sum_{m}^{i=1}\alpha_i-\omega^{\mathrm{T}}\sum_{i=1}^{m}\alpha_i x_i y_i$
$=\sum_{m}^{i=1}\alpha_i-\frac{1}{2}(\sum_{i=1}^{m}\alpha_i x_i y_i)^{\mathrm{T}}(\sum_{i=1}^{m}\alpha_i x_i y_i)$
$=\sum_{m}^{i=1}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j x_i^{\mathrm{T}}x_j y_i y_j$
$\alpha_i\geq 0,\sum_{i=1}^{m}\alpha_i y_i=0$

3. KKT 条件

- 为什么转换成对偶问题：
  1. 对偶问题将原始问题中的约束转为了对偶问题中的等式约束
  2. 方便核函数的引入
  3. 改变了问题的复杂度。由求特征向量 ω 转化为求比例系数 α，在原始问题下，求解的复杂度与样本的维度有关，即 ω 的维度。在对偶问题下，只与样本数量有关。

- KKT 条件：
  1. $\alpha_i\geq 0$
  2. $y_i(\omega^{\mathrm{T}}x+b-1)\geq 0$
  3. $\sum\alpha_i(y_i(\omega^{\mathrm{T}}x+b-1))=0$

4. 核函数

- 高斯核为什么有效?
  在现实任务中，原始样本空间内也许并不存在一个能正确划分两类样本的超平面。对这样的问题，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。令 $\phi(x)$ 表示将 x 映射后的特征向量，于是，在特征空间中划分超平面所对应的模型可表示为 $f(x)=\omega^{\mathrm{T}}\phi(x)+b$
- 常用的核函数
  1. 线性核
  2. 多项式核
  3. 高斯核
  4. 拉普拉斯核
  5.Sigmoid 核

5. 过拟合

- 松弛变量 ξ
  约束条件变为:$s.t. y_i(\omega^{\mathrm{T}}x+b)\geq 1-\xi_i$
  引入松弛变量使 SVM 能够容忍异常点的存在。为什么？因为引入松弛变量后，所有点到超平面的距离约束不需要大于等于 1 了，而是大于 0.8 就行了（如果 =0.2 的话），那么异常点就可以不是支持向量了，它就作为一个普通的点存在，我们的支持向量和超平面都不会受到它的影响。
- 软间隔支持向量机、、在最大化间隔的同时，不满足约束的样本应尽可能少。

$$min_{\omega,b,\xi_i}\frac{1}{2}||w||^2+C\sum_{i=1}^{m}\xi_i$$