1 风险函数

1.1 经验风险最小化 (empirical risk minimization,ERM)

极大似然估计 (maximum liklihood estimation) 就是经验风险最小化的一个例子。(容易过拟合，不加正则化)

1.2 结构风险最小化 (structural risk minimization,SRM)

贝叶斯估计中的最大后延概率估计 (maximum posterior probability estimation,MAP) 就是结构风险最小化的一个例子。(加了正则化)

2 生成模型，判别模型

生成模型就是生成 (数据的分布) 的模型。

判别模型就是判别 (数据输出) 的模型。

更进一步，从结果角度，两种模型都能给你输出量 (label 或 y etc.)。但，生成模型的处理过程会告诉你关于数据的一些统计信

息（p(x|y) 分布 etc.），更接近于统计学；而判别模型则是通过一系列处理得到结果，这个结果可能是概率的或不是，这个并不改变他是不是判别的。

3 感知机

3.1 关于感知机学习算法的对偶形式

我们假设样本点 (x_i, y_i) 在更新的过程中使用了 n_i 次。因此，从原始形式的学习过程中可得到，最后学习到的 w 和 b 可以分别表示为:

$$w = \sum_{i=1}^{N} n_i \eta y_i x_i \tag{1}$$

$$b = \sum_{i=1}^{N} n_i \eta y_i \tag{2}$$

考虑 n_i 的含义: 如果 n_i 的值越大，那么意味着这个样本点经常被误分。什么样的样本点容易被误分？很明显就是离超平面近的点。超平面稍微一动一点点，这个点就从正

变为负或者从负变正。如果学过 SVM 就会发现，这种点很可能就是支持向量。　　代入式 (1) 和式子 (2) 到原始形式的感知机模型中，可得:

$$f(x) = sign(w.x+b) = sign(\sum_{j=1}^{N} n_j \eta y_j x_j.x + \sum_{j=1}^{N} n_j \eta y_j) \tag{3}$$

此时，学习的目标就不再是 w 和 b, 而是 n_i,i=1,2,....N。

相应的，训练过程变成:

1. 初始时 $\forall n_i$=0。
2. 在训练集中选取数据 (x_i, y_i)。
3. 如果 $y_i(\Sigma_{j=1}^{N} n_j \eta y_j x_j.x_i + \Sigma_{j=1}^{N} n_j \eta y_j) \leq 0$, 更新:$n_i \Leftarrow n_i + 1$ 。
4. 转至 2 直至没有误分类数据。

可以看出，其实对偶形式和原始形式没有本质区别，但是从式 (3) 可以看出，样本点的特征向量以内积的形式存在于感知机对偶形式的训练算法中，因此，如果事先计算好所

有的内积，也就是 Gram 矩阵，就可以大大提高计算速度。

Gram 矩阵:

$$G = [x_i.x_j]_{N*N}$$

4 k 近邻

4.1 kd 树

5 朴素贝叶斯

5.1 先验概率

事件发生前的预判概率，可以是基于历史的数据统计，也可以由常识北京得出，也可以是人的主观观点给出，一半都是单独事件概率，如 P(x)。

5.2 后验概率

事件发生后求得反向条件概率 P(y|x)

5.3 贝叶斯公式

P(y|x) = (P(x|y) * P(y)) / P(x)

这里:

P(y|x) 是后验概率，一般是我们求解的目标。

P(x|y) 是条件概率，又叫似然概率，一般是通过历史数据统计得到。一般不把它叫做先验概率，但从定义上也符合先验定义。

P(y) 是先验概率，一般都是人主观给出的。贝叶斯中的先验概率一般特指它。

P(x) 其实也是先验概率，只是在贝叶斯的很多应用中不重要（因为只要最大后验不求绝对值），需要时往往用全概率公式计算得到。

5.4 最大似然理论

认为 P(x|y) 最大的类别 y 买就是当前文档所属类别