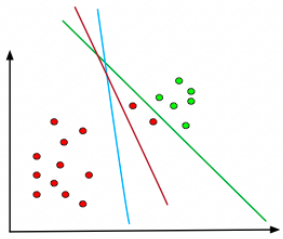
一、分类器的基本内容以及各个模型特点分析比较

1. 感知机
   - 感知机的主要工作就是用一个超平面将标签 Y 完全分割成两种特征 {+1, -1}
   - 其几何表达是：$f(x) = sign(\vec{w}\vec{x} + b)$，其中有：$sign(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}$
   - 感知机模型通过不断调整参数 w 和 b 使得完全分类成两种特征{+1, -1}
   - 相对于朴素贝叶斯来说，其优点是能够更加准确的分类成两部分，因为其分类目标就是完全分类成两部分。而朴素贝叶斯会对特征指定一种分布（如高斯分布），这就导致，其仍有一定概率被错分类（概率比较小的地方），相对于感知机，会降低一部分正确率。
   - 缺点是对于有一些数据集，感知机会追求最大程度完全划分，这就有可能造成过拟合。相比，与之模型相近的 SVM 在构造目标函数的时候，会给 margin 一定的 tradeoff（软间隔分类），它只追求大致的精确分类，一定程度上能够避免过拟合。而且，SVM 给出的是唯一的超平面，而感知机给出的超平面会随着初值变化而最终变化。
2. 朴素贝叶斯
   - 通过特征 X 来预测标签 Y 的概率：$\frac{P(\vec{x}|y)P(y)}{P(x)}$，其中$P(y = c) = \frac{\sum_1^n I(y_i = c)}{n}$
   - 假设每个特征x_α都是独立的，我们可以有：$P(\vec{x}|y) = \prod_1^d P(x_\alpha|y)$
   - 最大特点是对于每一个特征，我们都假设了其间是条件独立的。是一个生成模型。
   - 基于此特征，一个很大的缺点就是在关联性比较强的样本上，分类的结果就比较差。
   - 而其显著优点就是，因为基于独立假设，概率计算被大大简化，节省内存和时间。
3. 逻辑回归
   - 与高斯朴素贝叶斯对应的判别模型，建模形式为：$P(y|x_i) = \frac{1}{1+e\text{^}(-y(W_i^T x_i + b))}$
   - 添加一个参数$\vec{w}$，用 MLE（$P(y|\vec{x}, \vec{w})$）或者 MAP（$P(y|\vec{x}, \vec{w})P(\vec{w})$），找最优参数。
   - 缺点是很容易过拟合，需要训练大量数据，收敛速度相对于朴素贝叶斯来讲较慢。
   - 其优点就是，没有像朴素贝叶斯那样对于$P(x|y)$作出任何假设，所以分类更加灵活。但是如果特征分布提出的正确合理，朴素贝叶斯会收敛更快并得到和逻辑回归差不多的准确结果。
4. 支持向量机
   - 软间隔分类目标：$\min \frac{1}{2}||w||^2 + C\sum \max(1 - y_i(w^T x_i + b), 0)\ \ s.t. y_i(w^T x + b) \geq 1 - \max(1 - y_i(w^T x_i + b), 0)$ 。构造相应拉格朗日函数解决这个问题即可。
   - 硬间隔分类目标：去掉上述目标方程与约束条件中的松弛项即可。
   - 并不是所有线型分类数据集都是能完全线型可分的，软间隔中给出了松弛项，允许一定错误，防止过拟合。对于完全不可线型分类的数据我们还可以用核方法转换。
   - 这个模型与感知机模型非常相近，前面已经给出描述。主要特点就是，支持向量机给出的超平面是唯一的，而感知机的超平面会随着初值而变化，并且容易过拟合。
   - SVM 的优点就是在追求解决感知机过拟合问题的同时，也能够做到更加精确的分类，这使得 SVM 在大多数数据集中的效果表现更好。

二、综合比较说明

1. 感知机和朴素贝叶斯、感知机和 SVM 的比较

   关于感知机、朴素贝叶斯分类、SVM 分类（线性情况）时的特征如下图所示：



（1）绿线感知机、蓝线贝叶斯

- 可以发现，感知机分类结果更绝对，力求将两类点完全分开。而贝叶斯就将两类点按照高斯分布的状态分成两类，少数红点靠近绿点的部分被认为是概率较小的部分。（有的时候感知机的绝对会带来过拟合）

（2）绿线感知机、红线 SVM

- 同样的，在感知机和 SVM 的比较中，也会看出感知机相对于 SVM 来比有过拟合的问题。而 SVM 给出了 tardeoff，允许在 margin 内有点，一定程度上避免了感知机带来的过拟合。感知机则相对绝对。
- 进一步说明，感知机给出的超平面不是唯一的（不一定是绿线），他会随着初值的变化而最终变化，而 SVM 给出的超平面则是唯一的。

（3）综合比较

- 综合比较下来，感知机模型分类最绝对（但也可能导致过拟合）。贝叶斯的分类会使得分类比较类似于高斯分布的概率，减少了一定的过拟合，但也有可能导致不准确。而 SVM，这个模型指导我们选择了好的 tradeoff 值（即允许 margin 内有点的程度），其效果在绝大多树情况下表现都比较优秀（即一定程度上解决了过拟合，且保证了准确率）。

2. 朴素贝叶斯和逻辑回归比较的进一步说明

- 朴素贝叶斯中，对于每一个标签 y，都为其建模$P(x|y)$，而逻辑回归中直接对$P(y|x)$建模，没有对$P(x|y)$作出过多假设。
- 逻辑回归没有指定$P(x|y)$是否为独立分布，所以更加灵活，对于关联度大的特征集合，更准确；但如果朴素贝叶斯中提出的分布很合理，就没什么差别。
- 逻辑回归容易过拟合，需要训练更多的数据。
- 朴素贝叶斯一般比逻辑回归收敛更快。