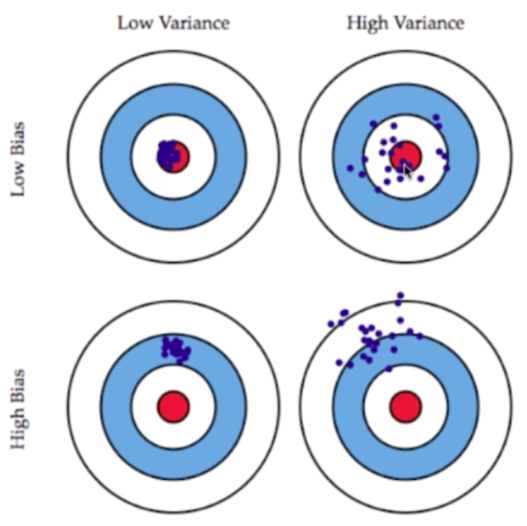
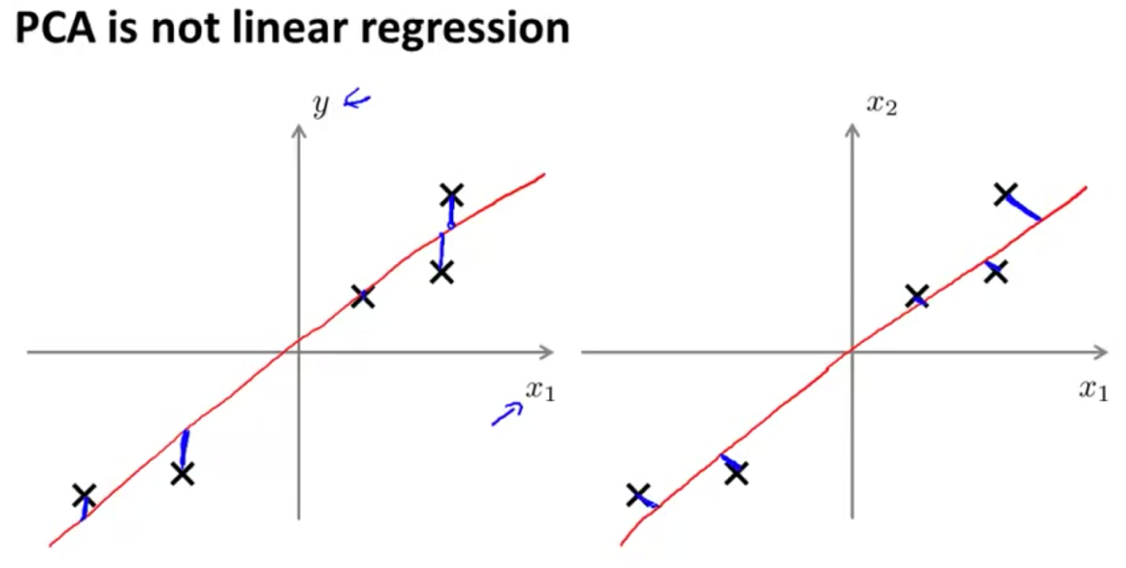
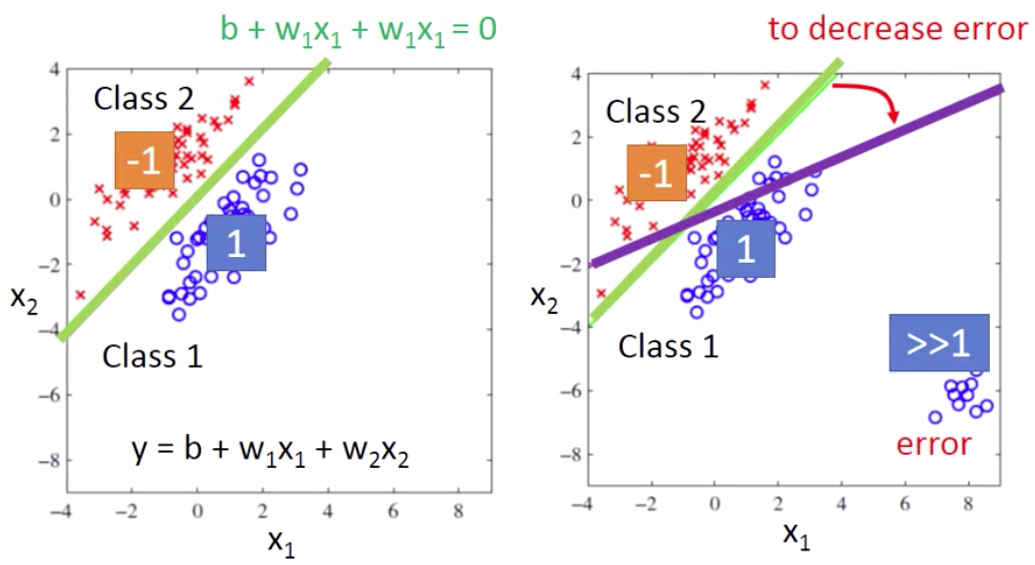
* 监督学习：有训练集，回归、分类。
* 非监督学习：无训练集，聚类。
* 半监督学习：一部分数据有标记，另一部分无标记；通常先使用无监督学习进行数据处理，再使用监督学习进行训练和预测。
* 强化学习：根据周围环境情况采取行动，根据采取行动的结果，学习行动方式。
* 模型参数：算法过程中学习的参数。
* 超参数：在算法运行前需要决定的参数，网格搜索寻找最优解。
* 参数学习算法：简单线性回归、逻辑回归、神经网络、朴素贝叶斯。
* 非参数学习算法： KNN、复杂线性回归、SVM、决策树。
* 训练模型：60%训练集(train)+20%验证集(cv)+20%测试集(test)；验证集调整超参数，测试集衡量最终模型性能。
* 拟合问题：欠拟合：模型简单，train、test错误大，高偏差；过拟合：模型复杂，train错误小，test错误大，高方差。
* 偏差、方差：简单模型高偏差低方差，复杂模型低偏差高方差。



* 解决欠拟合：增加模型复杂度、增加样本特征、减小正则化系数。
* 解决过拟合：减少模型复杂度、减少样本特征、增大正则化系数、增加训练集。
* 归一化：(1)min-max标准化：，归一化后x∈[-1,1]；，归一化后x∈[0,1]；多用于不涉及距离度量、协方差计算、数据不符合正态分布时，如图像处理RGB[0,255]；(2)0均值标准化：，归一化后x均值为0，标准差为1；多用于分类、聚类、PCA降维。要用训练集进行拟合，再将训练集和测试集归一化，而不是用测试集拟合后归一化测试集。
* 梯度下降法：，多次同时更新参数直到找到最优解；需要归一化；学习率过小导致求解过慢，学习率过大导致跳过最优解甚至无法收敛；越接近最优解的位置参数更新幅度越小，速度越慢；注意观察迭代次数和梯度值的图像，只要学习率足够小，梯度就不会出现上升现象；为避免只找到局部最优解，要多次运行、随机化初始点；可以训练具有大量特征(n>1w)的数据。
* 标准方程法：一次计算即可得到最优解；不需要归一化；时间复杂度O(n3)，不适合具有大量特征(n>1w)的数据；只适用于线性模型。
* 共轭梯度、L-BFGS、BFGS算法：不需要手动选择学习率，速度更快、更加复杂。
* 最小化损失函数J(θ)：在训练集中，利用梯度下降法调节θ把预测值与真实值的差降到最小。
* KNN：默认多分类、回归，性能强，效率低，高度数据相关，不具有可解释性，维数灾难；调参k(临近数)、weights(距离权重)、p(明可夫斯基距离)；；可使用KD-TREE、BALL-TREE优化。
* 线性回归：，最小二乘法，后续计算需要求导取极值，所以用平方不用绝对值。
* 一元线性回归：回归方程h(x)=θ0+ θ1x1。
* 多元线性回归：回归方程h(x)=θ0+θ1x1+θ2x2+θ3x3+……
* 多项式回归：基于线性回归模型，只是增加了新的高阶特征；回归方程h(x) = θ0+θ1x+θ2x2+θ3x3+……
* 线性回归正则化： J(θ)=，λ为正则化参数，调节λ解决过拟合和欠拟合，λ过大欠拟合，λ过小过拟合。
* L1正则化、L2正则化、弹性网：(1)L1正则化，LASSO Regression，，趋于直线，准确率低，效率高，可进行特征选择；(2)L2正则化，Ridge Regression，，趋于曲线，准确率高，效率低；(3)弹性网，Elastic Net，，准确率适中，效率适中。
* scikit-learn正则化：C·J(θ)+L1、C·J(θ)+L2，默认L2正则化。
* PCA：非监督学习，只用在训练数据集中，最小化投影误差，使用PCA前需归一化(0均值标准化)以减小误差，多用于降维、降噪、提取特征、压缩数据、减少内存和硬盘需求、处理图像数据；减小特征维度，一般保留95%-99%的数据，选择最小k：；不要用PCA解决过拟合问题。
* PCA与线性回归：



* 为什么不用回归思想处理分类问题：



* 逻辑回归：判别方程 (z=θ0+θ1x1+θ2x2+θ3x3......)，由于sigmoid函数的存在，使用平方差损失函数(y-h(z))2不能保证是凸函数，可能只能得到局部最优解，所以使用交叉熵损失函数，，使用最大似然估计，默认线性二分类，可以实现多分类。
* 混淆矩阵(Confusion Matrix)：真实值与预测值相同为T。

|  |  |  |
| --- | --- | --- |
| 预测值(N/P)  真实值 | 0(N) | 1(P) |
| 0 | 预测negative正确  TN | 预测positive错误  FP |
| 1 | 预测negative错误  FN | 预测positive正确  TP |

* 精确率(P)：预测positive正确/positive总数，TP/(TP+FP)。
* 召回率(R)：预测positive正确/真实总数，TP/(TP+FN)。
* F-Score：，综合查准率和召回率，越高越好。
* PR曲线：x召回率，y精准率，曲线左下方面积越大模型越好。
* 真正类率(TPR)： 预测的正类中实际正实例占所有正实例的比例，TP/(TP+FN)。
* 假正类率(FPR)：预测的正类中实际负实例占所有负实例的比例，FP/(FP+TN)。
* ROC曲线：x假正例率，y真正例率，曲线右下方面积越大模型越好。
* 灵敏度：TP/(TP+FN)。
* 特异度：TN/(FP+TN)。
* PR、ROC比较：正负样本数量差距不大时，PR与ROC趋势相似，但负样本较多时，ROC效果看似较好，PR效果却很差，因为PR更关注正样本，PR在正负样本比例悬殊较大时更能反映分类的性能。
* 支持向量机：，默认二分类，C过大容易过拟合，C过小容易欠拟合。
* 核函数：线性核函数主要用于线性可分样本；多项式核函数可以实现低维映射到高维，参数多计算量大，容易过拟合；RBF(高斯)核函数可以实现低维映射到高维，一般使用欧式距离，应用最广泛，，，σ2过大，gamma过小，函数较平滑，产生欠拟合；σ2过小，gamma过大，函数较陡峭，产生过拟合。
* 逻辑回归VS支持向量机：n=特征数量，m=样本数量，(1)n>>m：n=10000，m=10-1000，使用逻辑回归或无核(线性核)函数SVM；(2)n<<m，n=1-1000，m=50000+，手动增加特征之后，使用逻辑回归或无核(线性核)函数SVM；(3)n、m适中，n=1-1000，m=10-10000，使用高斯核函数SVM。
* 决策树：默认多分类、回归，具有可解释性，每一个叶子结点是一个种类，容易过拟合，对个别样本很敏感，只能划分垂直于坐标轴的边界，十分依赖于调参，很少单独使用。
* 决策树构造方法：(1)信息熵，，pi是每一类信息所占的比例，信息熵越大不确定性越高，信息熵最小时进行类别划分；(2)基尼系数，，pi是每一类信息所占的比例，基尼系数越大不确定越高，基尼系数最小时进行类别划分；信息熵计算稍慢，基尼系数计算稍快，二者没有明显优劣。
* 集成学习：(1)Voting，多个模型投票，hard少数服从多数，soft加权投票(需要计算每个模型概率)；(2)Bagging，每个子模型只处理一部分数据，每个子模型不需要太高的准确率，集成学习后准确率很高；可对样本和特征进行放回取样，平均有37%的样本取不到(OOB-Out of bag)，可以不区分训练集核测试集；首选决策树模型。(3)Boosting，AdaBoosting(提升树)每个模型都尝试增强整体的效果,GradientBoosting(GDBT)每训练一个新模型会根据前一个模型的错误进行改进。(4)Stacking：类似神经网络，把训练集分成若干部分，建立若干层，每层有若干个训练模型，把每部分训练数据输入对应层，前一层的输出作为下一层的输入。
* Bagging和Boosting区别：(1)样本选择：Bagging训练集是在原始集中有放回选取的，从原始集中选出的各轮训练集之间是独立的；Boosting：每一轮的训练集不变，只是训练集中每个样例在分类器中的权重发生变化，权值是根据上一轮的分类结果进行调整。(2)样例权重：Bagging使用均匀取样，每个样例的权重相等；Boosting根据错误率不断调整样例的权值，错误率越大则权重越大。(3)预测函数：Bagging所有预测函数的权重相等；Boosting每个弱分类器都有相应的权重，对于分类误差小的分类器会有更大的权重。(4)并行计算：Bagging各个预测函数可以并行生成；Boosting各个预测函数只能顺序生成，因为后一个模型参数需要前一轮模型的结果。
* 集成学习与决策树关系：(1)Bagging + 决策树 = 随机森林；(2)AdaBoost + 决策树 = 提升树；(3)Gradient Boosting + 决策树 = GBDT。
* 随机森林：决策树在节点划分上，在随机的特征子集上寻找最优划分特征。
* 极其随机森林：决策树在节点划分上，使用随机的特征和随机的阈值；提供额外的随机性，抑制过了拟合，但增大了偏差，训练速度更快。
* 朴素贝叶斯：条件独立，主要应用于文本分类，准确率较高，训练集的准确性影响较大。
* K-means：确定类别个数n，随机取n个点作为每个族群中心点，计算其余点到中心点额度距离，按照最近距离分成三族群，分别计算三个族群平均值作为族群新中心点，与旧中心点比较，反复操作直至新中心点与旧中心点重合结束聚类；，非监督学习，聚类，K取值2-10；轮廓系数评估模型，bi外部距离，ai内部距离。
* 代码：大写字母X表示data矩阵，小写字母y表示label向量；表示第i个样本的第j个特征值；网格搜索寻找最佳超参数GridSearchCV；预处理中多项式在标准化前，标准化在PCA前。