1、

逻辑回归相比线性回归，有何异同?

逻辑回归和线性回归之间既有区别又有联系。逻辑回归和线性回归最大的不同点是逻辑回归解决的是分类问题而线性回归则解决的是回归问题。逻辑回归又可以认为是广义线性回归的一种特殊形式，其特殊之处在于其目标(label/target)的取值服从二元分布。

所谓逻辑回归是一种特殊的广义线性回归，我们可以通过狭义线性回归到逻辑回归的转化过程来理解。

狭义线性回归的表达式可表示为:

$$y = w * x + b$$

如果我们希望这个模型可以对二分类任务做出预测，即目标满足 0,1 分布。那么希望预测出来的值经过某种转换之后，大部分可以分布在 0,1 两个值附近。

我们发现 sigmoid 函数可以帮助我们做这样的转换，sigmoid 函数的表达式为

$$\delta(z) = \frac{1}{1 + e^{-z}}$$

令:

$$y = \delta(z)$$

则:

$$\log\frac{y}{1-y} = z = w * x + b$$

可以看到相比于狭义线性回归采用 y 作为预测结果，逻辑回归则采用$\log\frac{y}{1-y}$作为预测结果。逻辑回归还可以表示为:

$$y = sigmoid(w * x + b)$$

通过以上的步骤推演，我们知道逻辑回归的求值计算其实就是在线性回归的基础上，再做一个 sigmoid 计算。所以它们都可以用相同的方法比如梯度下降来求解参数。

回归问题常用的性能度量指标?

-------- 点对点误差 --------

1. **MSE(Mean Square Error) 均方误差** – 该统计参数是预测数据和原始数据对应误差的平方和的均值

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2. **RMSE (Root Mean Square Error) –** 观测值与真值偏差的平方和与观测次数 n 比值的平方根，用来衡量观测值同真值之间的偏差

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

3. **MAE (Mean Absolute Error) –** 计算模型输出与真实值之间的平均绝对误差

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}_i|$$

------ 带归一化的误差求解方法 ------

4. **MAPE (Mean Absolute Percentage Error) –** 不仅考虑预测值与真实值误差，还考虑误差与真实值之间的比例

$$MAPE = \frac{1}{n}\sum_{i=0}^{n}\frac{|y_i - \hat{y}_i|}{y_i}$$

5. **MASE (Mean Absolute Scaled Error) –** 平均平方百分比误差

$$MASE = \frac{1}{n}\sum_{i=0}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2$$

------- 点对面误差 ----------

6. **R-Square 决定系数($coeffcient\ of\ determination$)**

$$R - squared = 1 - \frac{RSS}{TSS}$$

其中 RSS (residual sum of squares)的表达式为：

$$RSS = \sum_{i}^{n}(y_i - \hat{y}_i)^2$$

TSS (total sum of squares) 的表达式为：

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

常见分类问题的度量指标?

实际类别	预测类别			
		Yes	No	总计
	Yes	TP	FN	P（实际为Yes）
	No	FP	TN	N（实际为No）
	总计	P'（被分为Yes）	N'（被分为No）	P+N

准确率-所有预测正确的样本（正样本预测为正，负样本预测为负）与所有样本的比值。

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

精确率-**精确率**是针对我们**预测结果**而言的，它表示的是预测为正的样本中有多少是真正的正样本。那么预测为正就有两种可能了，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)，也就是

$$Precision = \frac{TP}{TP + FP}$$

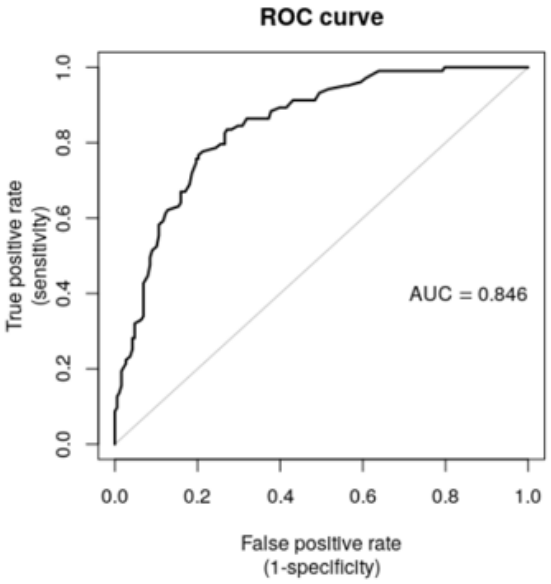
召回率-**召回率**是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类(FN)。

$$Recall = \frac{TP}{TP + FN}$$

F1 值-**F1 值**是精确率和召回率的调和值

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

AUC (Area Under Curve), 曲线下的面积。这里的 Curve 指的的是 ROC (receiver operating characteristic curve) 接收者操作特征曲线，是反映敏感性(sensitivity)和特异性(1-specificity)连续变量的综合指标, ROC 曲线上每个点反映着对同一信号刺激的感受性。ROC 曲线是通过取不同的阈值来分别计算在每个阈值下，伪正类率 FPR（False Positive Rate）和真正类率 TPR （True Positive Rate）的值来绘制的。



分类任务的极大似然损失函数 – 参见题目（逻辑回归的损失函数？）

逻辑回归的损失函数？

逻辑回归的预测结果服从 0, 1 二项分布，假设预测为 1 的概率为 p, 则

$$P(Y=1|x)=p(x), P(Y=0|x)=1-p(x)$$

写出似然函数来为：

$$L(w)=\prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

其中 n 为样本数，y_i 是 label 取值为 0 或 1。对上面的似然函数两边求对数可得

$$logL(w)=\sum_{i=1}^{n}[y_i logp(x_i)+(1-y_i)\log(1-p(x_i))]$$

极大似然估计方法就是要求函数参数使得 logL(w)最大，所以我们也可以将目标定为求函数参数使得-logL(w)最小。取平均之后，就得到最终的目标函数为：

$$J(w)=-\frac{1}{n}logL(w)=-\frac{1}{n}\sum_{i=1}^{n}[y_i logp(x_i)+(1-y_i)\log(1-p(x_i))]$$

极大似然估计 是一种根据结果即 预测值来推测参数的一种普遍方法。提到极大似然是为了引出逻辑回归因为二项分布而得到的似然函数

逻辑回归处理多标签分类问题时，一般怎么做？

如果 y 不是在[0,1]中取值，而是在 K 个类别中取值，这时问题就变为一个多分类问题。有两种方式可以处理该类问题：

当 K 个类别不是互斥的时候，即每次对样本进行分类时，不需要考虑它是不是还可能是别的类别；那么我们可以为每个类别建立一个逻辑回归模型。用它来判断样本是否属于当前类别。

当 K 个类别是互斥的时候，即 y=i 的时候意味着 y 不能取其他的值，这种情况下 Softmax 回归更合适一些。Softmax 回归是直接对逻辑回归在多分类的推广，相应的模型也可以叫做多元逻辑回归（Multinomial Logistic Regression）。模型通过 softmax 函数来对概率建模，具体形式如下：

$$P = (y = i|x, \theta) = \frac{e^{\theta_i x}}{\sum_j^K e^{\theta_j x}}$$

决策函数为：

$$y^* = argmax_i P(y = i|x, \theta)$$

对应的损失函数为：

$$J(\theta) = -\frac{1}{N}\sum_i^N \sum_j^K 1(y_i = j) log \frac{e^{\theta_i x}}{\sum_k^K e^{\theta_k x}}$$