Introduction:
Empiricism - Positivism - Logical positivism - Verificationism - Pragmatism - Falsificationism

How to make a decision:
• The classical statistics: Significance testing.
a tool for decision making under uncertainty

Null hypothesis (H0) (no effect) and its complement (H1).
Significance level (α, p value).
Statistical power (1- β), probability of rejecting H0 given that it should be rejected.
Sample size n / Effect size.

Power curve: central limit theorem, whether under the curve,
(Assume complete ignorance about the world ("anything possible").
Formulate a hypothesis H0 and compare P(data|H0) to a pre-defined probability threshold (significance level) α.
If P(data|H0)<α, reject H0 (the data are unlikely/surprising if H0 is true).
(If the observed mean falls outside the 95% range, we reject H0))

risks
Type I errors (reject H0 when true).
Type II errors (accept H0 when false).
Whether the "correct" significance level? α = 0.05

Decision making: we need to consider uncertainty and preferences.
Probability is a measure of uncertainty
Utility is a measure of preference that combines with probability as mathematical expectation.
The main problem (of course, after determining what the value of the outcomes are) is to determine the **prior probability of the hypothesis**. To see that, start with Pr(H0|D) and the derive everything using Bayes theorem in terms of Pr(H0), Pr(D|H0), and Pr(D|H1).
Recall the possible errors are: (type I and type II) - α and β are probabilities of these errors.

Decision-theoretic (Bayesian) view allows to explore the exact relation between the significance level and the decision.

Confidence Intervals

from the point of view of the **classical hypothesis testing**

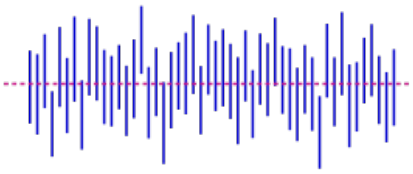
I'm 95% sure that the true mean is inside this interval

标准差：standard deviation

Standard error: $\frac{\sigma}{\sqrt{n}} = 0.5 \text{ grams}$

$$0.95 = 1 - \alpha = P(-z \le Z \le z) = P\left(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right)$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$



Confusion matrix: 混淆矩阵

Sensitivity: recall rate, measures the proportion of actual positives which are correctly identified

Specificity: measures the proportion of negatives which are correctly identified

A perfect predictor would be described as 100% senitivity and 100% specificity.

The fecal occult blood (FOB) screen test used in 2,030 people to look for bowel cancer

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)	Positive predictive value = Σ True Positive / Σ Test Outcome Positive
	Test Outcome Negative	False Negative (Type II error)	True Negative	Negative predictive value = Σ True Negative / Σ Test Outcome Negative
		Sensitivity = Σ True Positive / Σ Condition Positive	Specificity = Σ True Negative / Σ Condition Negative	

http://en.wikipedia.org/wiki/Specificity_(statistics)

Accuracy=(TN+TP)/(TN+TP+FN+FP)=1840/2030=90.6%

Related calculations

False positive rate (α) = type I error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%

False negative rate (β) = type II error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) = 33%

Power = sensitivity = 1 − β

Likelihood ratio positive = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4

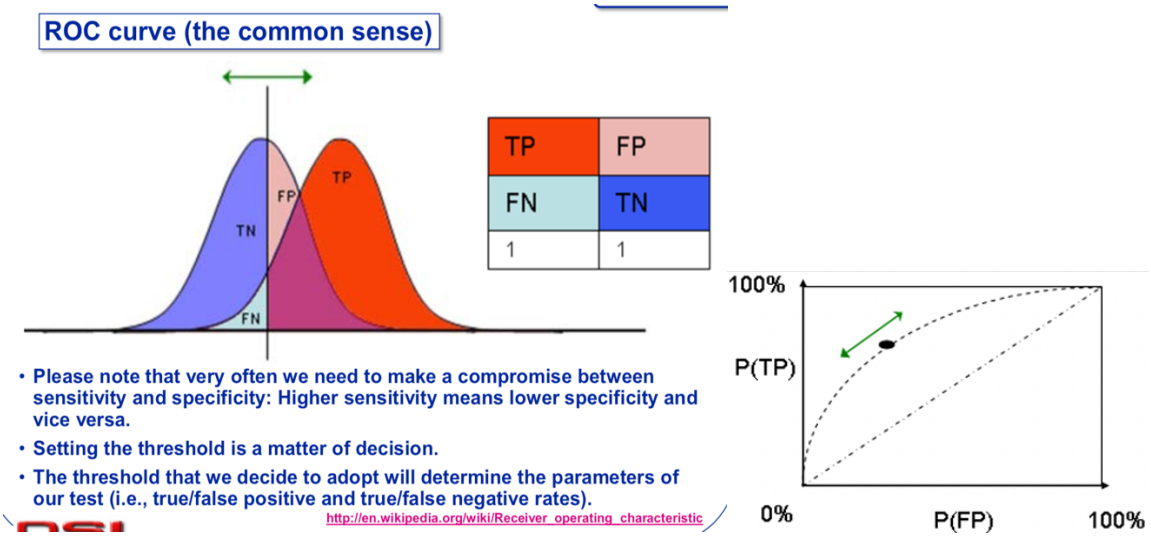
Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = 10%
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ 99.5%
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ 67%	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = 91%	

Hence, with large numbers of false positives and few false negatives, a positive FOB screen test is in itself poor at confirming cancer (PPV = 10%) and further investigations must be undertaken; it did, however, correctly identify 66.7% of all cancers (the sensitivity). However as a screening test, a negative result is very good at reassuring that a patient does not have cancer (NPV = 99.5%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

http://en.wikipedia.org/wiki/Specificity_(statistics)

ROC (Receiver Operating Characteristic) Curves



对于某个**二分类**分类器来说，输出结果标签（0 还是 1）往往取决于输出的概率以及预定的概率阈值，比如常见的阈值就是 0.5，大于 0.5 的认为是正样本，小于 0.5 的认为是负样本。如果**增大**这个阈值，预测错误（针对正样本而言，即指预测是正样本但是预测错误，下同）的概率就会降低但是随之而来的就是预测正确的概率也降低；如果**减小**这个阈值，那么预测正确的概率会升高但是同时预测错误的概率也会升高。实际上，这种阈值的选取也一定程度上反映了分类器的**分类能力**。我们当然希望无论选取多大的阈值，分类都能尽可能地正确，也就是希望该分类器的分类能力越强越好，一定程度上可以理解成一种**鲁棒能力**吧。

- 假阳率，简单通俗来理解就是预测为正样本但是预测错了的可能性，显然，我们不希望该指标太高。

$$FPR = \frac{FP}{TN + FP}$$

- 真阳率，则是代表预测为正样本但是预测对了的可能性，当然，我们希望真阳率越高越好。

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

显然，ROC曲线的横纵坐标都在[0,1]之间，自然ROC曲线的面积不大于1。现在我们来分析几个**特殊情况**，从而更好地掌握**ROC曲线的性质**：

- (0,0)：假阳率和真阳率都为0，即分类器全部预测成负样本
- (0,1)：假阳率为0，真阳率为1，全部完美预测正确，happy
- (1,0)：假阳率为1，真阳率为0，全部完美预测错误，悲剧
- (1,1)：假阳率和真阳率都为1，即分类器全部预测成正样本
- TPR＝FPR，斜对角线，预测为正样本的结果一半是对的，一半是错的，代表随机分类器的预测效果

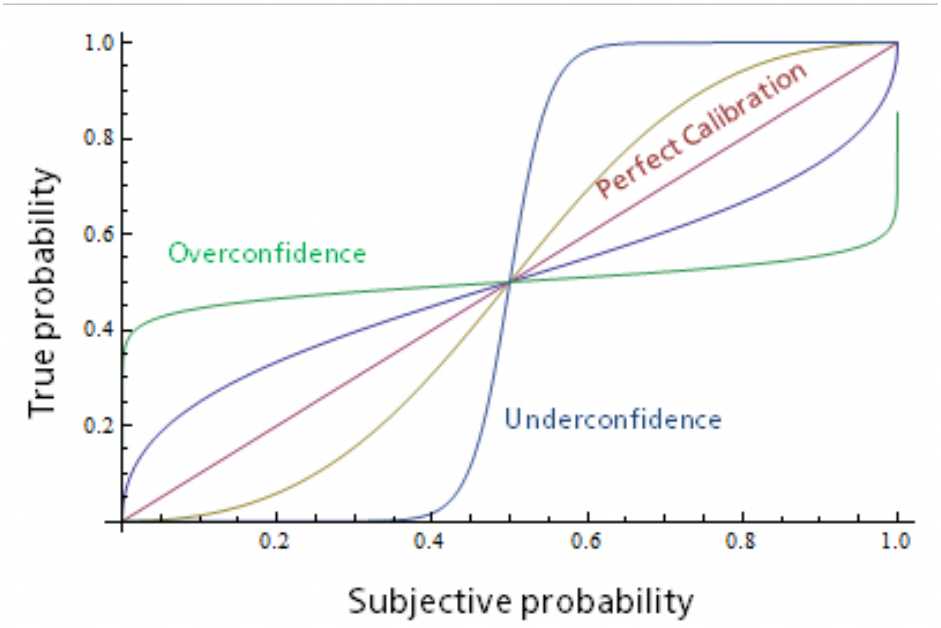
于是，我们可以得到基本的结论：**ROC曲线在斜对角线以下，则表示该分类器效果差于随机分类器，反之，效果好于随机分类器，当然，我们希望ROC曲线尽量除于斜对角线以上，也就是向左上角（0,1）凸。**

AUC: Area Under the (ROC) Curve

- AUC ＝ 1，代表完美分类器
- 0.5 < AUC < 1，优于随机分类器
- 0 < AUC < 0.5，差于随机分类器

Calibration

We plot the frequencies observed in the data (y axis) against the probabilities calculated by the system (x axis).



Cross-validation:

Testing a model on the same data that we used for training it does not seem fair.

It will favor most complex models that fit the data best.

Simpler model may actually fit future instances of data better than complex models.

Cross-validation prevents over-fitting

Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set.

It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

Divide the data into two disjoint sets: (1) training set and (2) test set (a.k.a. validation set).

Perform the analysis on the training set and validate the results on the test set.

Simple and effective.

Disad: It wastes data that could have been used for learning.

k-Fold cross-validation

"Leave-One-Out" cross-validation

Uses effectively n-1 instances for training and tests the model on all n instances, one at a time (an extreme case of k-fold, k=n).

Bootstrap cross-validation

Repetitive test set method, each repetition involves selecting a new test set from among all records.