Determining AUC from a score vector

nAgarAjan naTarAjan, vishvAs vAsuki

April 19, 2010

1 Problem

The problem is to identify the set of items with a certain property from given set of items.

2 Notation and Terminology

Let U be the universe of items, such that $|U| = u$. Let $T \subseteq U$ be the set of items having the property we are interested in.

2.1 Score generator

The task of a "score generator" is to output a score function $score(i)$, which can be used in generating a partial ordering of the items such that an item with a higher score is perceived to be more likely in T.

2.2 Predictor

Using the scores produced by a score generator, the associated predictor, parameterized by n identifies a set of n candidates called the *prediction*, P_n. The predictor works as follows.

Consider the bag of scores produced by the action of $score()$ on U. Let $sortedScores$ be a vector of these scores arranged in non-increasing order. The cutoff score is $c = sortedScores_n$.

Then, let $GT = \{i : score(i) > c\}$. Let $EQ = \{i : score(i) = c\}$. Let $C \subseteq EQ$ be a set of $n - |GT|$ distinct items selected uniformly at random from EQ. Then, $P_n = GT \cup C$.

Another way of looking at the action of a predictior by the following algorithm 1.

3 Evaluating a score generator at n

Suppose that the scores from a score generator are used in producing the prediction P_n.

Algorithm 1 Predictor

P = $\emptyset$
for i = 1 to n **do**
  Select an element k at random from $\{j : score(j) = sortedScores_i\}$.
  P = P $\cup \{k\}$.
end for

3.1 Sensitivity and Specificity

Sensitivity $X = \frac{|P_n \cap T|}{|T|}$, measures the ability of the predictor to identify items in T. Specificity $Y = \frac{|(U-P_n)\cap(U-T)|}{|U-T|}$, measures the ability of the predictor to exclude items not in T.

The problem is to find $E[X]$ and $E[Y]$, given $score()$.

3.2 Expected Sensitivity

For every $i \in T$, let X_i be a binary random variable, which is 1 if $i \in P_n$ and 0 otherwise. Then, $X = (1/|T|)\sum_i X_i$. By linearity of expection, $E[X] = (1/|T|)\sum_i E[X_i]$.

$$\begin{aligned}
Pr(X_i = 1|score(i) = c) &= \frac{n - |GT|}{|EQ|} \\
Pr(X_i = 1|score(i) > c) &= 1 \\
Pr(X_i = 1|score(i) < c) &= 0 \\
\forall i \in T \cap GT : E[X_i] &= 1 \\
\forall i \in T \cap EQ : E[X_i] &= \frac{n - |GT|}{|EQ|} \\
E[X] &= (1/|T|)(|GT \cap T| + |EQ \cap T|\frac{n - |GT|}{|EQ|})
\end{aligned}$$

Note that $EQ \geq n - |GT|$.

3.2.1 Sanity check

Consider what happens when $\forall i : score(i) = 0$. Then, $|GT| = 0, |EQ \cap T| = |T|, E[X] = \frac{n}{|EQ|} = \frac{n}{|U|}$, as expected.

3.3 Expected Specificity

$$\begin{aligned} Y &= \frac{|(U-P_n)\cap(U-T)|}{|U-T|} \\ &= \frac{|(U-T)-P_n\cap(U-T)|}{|U-T|} \\ &= \frac{|(U-T)|-|P_n\cap(U-T)|}{|U-T|} \\ &= 1-\frac{n-|T|X}{|U|-|T|} \\ &= 1-\frac{n-|T|X}{|U|-|T|} \\ E[Y] &= 1-\frac{n-|T|E[X]}{|U|-|T|} \end{aligned}$$

3.3.1 Sanity check

Consider what happens when $\forall i : score(i) = 0$. Then, $E[X] = \frac{n}{|U|}$, $E[Y] = 1 - \frac{n}{|U|} = \frac{|U|-n}{|U|}$, as expected.

3.4 Summary

In summary, the performance of a score generator at n is evaluated as follows. $|U|$ and $|T|$ will already be known. The cutoff score c is determined. The score vector is used to determine the sets GT, EQ. These are used to evaluate expected sensitivity and expected specificity.

4 Evaluating a score generator over the entire range of n

4.1 ROC and AUC

The sensitivity vs (1-specificity) plot for varying 'number of prediction parameters n is called ROC curve. Note that, both sensitivity and (1-specificity) are monotonically non-decreasing functions of n. The expected ROC curve, or E[ROC], can be produced by determining $E[X]$ and $E[Y]$ for various values of $n \in [1, |U|]$. The area under the ROC curve, AUC, is a measure of the overall performance of the score generator.

4.1.1 Evaluating expected AUC

E[AUC] can be approximated analytically with the area under a piecewise-linear function to approximate ROC. Alternatively, one can calculate E[AUC] exactly using score() and T as explained below.