Computational Intelligence: Methods and Applications

Lecture 16

Model evaluation and ROC

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Confusion matrices

For n observations, including n_+ form class ω_+ and n_- from other classes labeled ω_- prediction of a model M are counted and put in a confusion matrix (rows=true, columns=predicted by model M),

$$P(\omega \mid \omega_M) = \frac{1}{n}\left(\begin{array}{c|ccc|c} T\downarrow & \omega_+ & \omega_- & \omega_r & \leftarrow M \\ \omega_+ & n_{++} & n_{+-} & n_{+r} & n_+ \\ \omega_- & n_{-+} & n_{--} & n_{-r} & n_- \end{array}\right)$$

$$= \left(\begin{array}{c|ccc|c} & \omega_+ & \omega_- & \omega_r & \\ \omega_+ & P_{++} & P_{+-} & P_{+r} & P_+ \\ \omega_- & P_{-+} & P_{--} & P_{-r} & P_- \end{array}\right) = \left(\begin{array}{c|ccc|c} & \omega_+ & \omega_- & \omega_r & \\ \omega_+ & TP & FN & R_+ & P_+ \\ \omega_- & FP & TN & R_- & P_- \end{array}\right)$$

P_+ (P_-) is the priori probability of class ω_+ (ω_-) estimated from data.

FP and FN

Confusion matrix, including possibility of rejection (don't know answer):

$$P(\text{true} \mid \text{predicted}) = \begin{bmatrix} \text{TP} = P_{++} & \text{FN} = P_{+-} \\ \text{FP} = P_{-+} & \text{TN} = P_{--} \end{bmatrix}$$

This notation is used especially in medical applications:

P_{++} is a hit or true positive (TP); P_{++}/P_+ is a true positive rate;
P_{--} is a hit or true negative (TN); P_{--}/P_- is a true negative rate
P_{-+} false alarm, or false positive (FP); ex. healthy predicted as sick.
P_{+-} is a miss, or false negative (FN); ex. sick predicted as healthy.

Accuracy and errors

Models are frequently evaluated on the basis of their accuracy.
Elements of P_{ij} depend on the evaluated model, $P_{ij}(M)$

$$A(M) = P_{++}(M) + P_{--}(M)$$
$$L(M) = P_{-+}(M) + P_{+-}(M)$$
$$R(M) = P_{+r}(M) + P_{-r}(M) =$$
$$= 1 - L(M) - A(M)$$

Accuracy for class k and balanced accuracy used for unbalanced data:

$$A_k(M) = P_{kk}(M) / P_k$$
$$B(M) = \frac{1}{2}\left(A_+(M) + A_-(M)\right)$$

What is better?

Rank 4 classifiers producing the following confusion matrices:

$$P(\omega \mid \omega_{M1}) = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} \quad P(\omega \mid \omega_{M2}) = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}$$

$$P(\omega \mid \omega_{M3}) = \begin{bmatrix} 0.25 & 0.5 \\ 0.0 & 0.25 \end{bmatrix} \quad P(\omega \mid \omega_{M4}) = \begin{bmatrix} 0.4 & 0.0 \\ 0.2 & 0.4 \end{bmatrix}$$

Accuracy 0.5 Accuracy 0.8

But which one is better for given accuracy?
M1 or *M3*? *M2* or *M4*?

Ordering is meaningless because there are two parameters, not one! For example, accuracy for class 1 and class 2.

Other evaluation measures

In medical applications: accuracy for class + is called sensitivity: what percentage of really sick people does this test recognize?

Accuracy for class – is called specificity:
is this test specific to class + or does it always says + ?

In information retrieval sensitivity is called recall: what % of truly relevant information has been recalled?

Precision measures % of really relevant info among all recalled info.

F-measure is the harmonic mean of recall and precision.

$$S_+(M) = P_{+|+}(M) = P_{++}(M)/P_+$$

$$S_-(M) = P_{-|-}(M) = P_{--}(M)/P_-$$

$$\mathrm{Prec}(M) = P_{++}(M)/P_+(M) = P_{++}(M)/\left(P_{++}(M) + P_{-+}(M)\right)$$

$$F(M) = 2\mathrm{Recall}\cdot\mathrm{Prec}/\left(\mathrm{Recall}+\mathrm{Prec}\right)$$

Example

N=100 observations, for x=0 let 10 be from ω_+, and 30 from ω_- classes, and for x=1 there are 40 and 20 cases from ω_+, and ω_- classes:

$$N(\omega, x) = \begin{bmatrix} 10 & 40 \\ 30 & 20 \end{bmatrix}; \quad P(\omega, x) = \begin{bmatrix} 0.1 & 0.4 \\ 0.3 & 0.2 \end{bmatrix}$$

$$P(x=0) = 0.4; P(x=1) = 0.6; P(\omega_+) = P(\omega_-) = 0.5;$$

$$P(\omega \mid x) = P(\omega, x)/P(x) = \begin{bmatrix} 1/4 & 2/3 \\ 3/4 & 1/2 \end{bmatrix}$$

MAP rule for this data is:

if x=0 then select ω_- because $P(\omega_-|x=0) > P(\omega_+|x=0)$

if x=1 then select ω_+ because $P(\omega_+|x=1) > P(\omega_-|x=1)$

If x=0 then 30 ω_- vectors are assigned to ω_-, or $P(\omega_-,\omega_-)$=0.3 and 10 ω_+ vectors assigned to ω_- class, so $P(\omega_+,\omega_-)$=0.1; $P(\omega_+,\omega_+)$=0.2; $P(\omega_-,\omega_+)$=0.4

Example

If x=1 then 30 ω_+ vectors are assigned to ω_+, or $P(\omega_+,\omega_+)$=0.4 and 20 ω_- vectors assigned to ω_+ class, so $P(\omega_-,\omega_+)$=0.2

Therefore MAP decisions lead then to the confusion matrix:

with P_+=0.5, P_- =0.5 and
Accuracy = 0.4+0.3=0.7,
Error = 0.1+0.2=0.3

$$P\left(\omega_i \mid \omega_j(M)\right) = \begin{bmatrix} 0.4 & 0.1 \\ 0.2 & 0.3 \end{bmatrix}$$

Sensitivity =0.4/0.5=0.8 (recall)
Specificity =0.3/0.5=0.6
Balanced accuracy = (0.8+0.6)/2 = 0.7
Precision = 0.4/0.6 = 2/3 = 0.67
F-measure = 2*0.8*2/3 *1/(0.8+2/3) = 16/22 = 8/11=0.73

Error functions

Accuracy of the model M is usually maximized, it is a sum of TP+TN, or a combination of sensitivity and specificity with the class priors as weights:

$$A(M) = P_{++}(M) + P_{--}(M) = P_{+}S_{+}(M) + P_{-}S_{-}(M)$$

This is obvious: class + has P_+ fraction of all cases and sensitivity S_+ is the accuracy for this class, same for class –, so the sum gives accuracy.

Equivalently the error ratio may be minimized:

$$L(M) = 1 - A(M) = P_{+-}(M) + P_{-+}(M)$$

Sometimes we would like to make decisions if we are quite sure that they are correct. Confidence in model M may be increased by rejecting some cases. Error=sum of the off-diagonal confusion matrix $P(\omega_i, \omega_{Mj})$ elements (true vs. predicted), accuracy=sum (trace) of diagonal elements, so

$$E(M;\gamma) = \gamma L(M) - A(M) = \gamma \sum_{i \neq j} P(\omega_i, \omega_{Mj}) - \mathrm{Tr}\, P(\omega_i, \omega_{Mj}) \geq -1$$

More error functions

This combination for 2 classes (ex. one class against all the rest) is:

$$E(M;\gamma) = \gamma\left(P_{+-} + P_{-+}\right) - \left(P_{++} + P_{--}\right) \geq -1$$

Rejection rate, or the fraction of the samples that will not be classified is:

$$R(M) = 1 - L(M) - A(M)$$

Minimization of the error-accuracy is thus equivalent to error + rejection:

$$\min_M E(M;\gamma) \Leftrightarrow \min_M \left\{(1+\gamma)L(M) + R(M)\right\}$$

For $\gamma = 0$ this is equal to error + rejection; for large γ minimization of this error function over parameters of the model M reduces the sum of FP and FN errors, but at a cost of growing rejection rate; for example if the model M {$x > 5$ then ω_1 else ω_2} makes 10 errors, all for $x \in [5,7]$ then leaving samples in this range unclassified gives M' {$x > 7$ then ω_1 or $x \leq 5$ then ω_2}, no errors, but lower accuracy, higher rejection rate and high confidence.

Errors and costs

Optimization with explicit costs:
assume that FP (false alarms) cost α times more than FN (misses).

$$\begin{aligned}\min_M E(M;\alpha) &= \min_M \left\{P_{+-}(M) + \alpha P_{-+}(M)\right\} \\ &= \min_M \left\{P_{+}(1 - S_{+}(M)) - P_{+r}(M) + \right. \\ &\left. \alpha\left[P_{-}(1 - S_{-}(M)) - P_{-r}(M)\right]\right\}\end{aligned}$$

For α = 0 this is equivalent to maximization of

$$\min_M E(M;\alpha = 0) = \min\left\{P_{++}(M) + P_{+r}(M)\right\}$$

and for large α to the maximization of

$$\min_M E(M;\alpha \gg 1) = \min\left\{P_{--}(M) + P_{-r}(M)\right\}$$

Lifts and cumulative gains

Technique popular in marketing, where cumulative gains and "lifts" are graphically displayed: lift is a measure of how effective model predictions are = (results obtained with)/(without the predictive model).

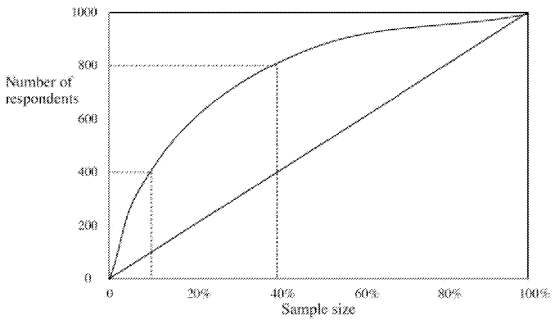
Ex: is X^i likely to respond? Should I send him an offer?
Suppose that 20% of people respond to your offer.
Sending this offer randomly to N people gives Y_0=0.2*N replies.

A predictive model (called "response model" in marketing), $P(\omega|X;M)$ uses information X to predicts who will respond.

Order predictions from the most likely to the least likely:
$P(\omega_+|X^1;M) > P(\omega_+|X^2;M) \ldots > P(\omega_+|X^k;M)$
The ideal model should put those 20% that will reply in front, so that the number of replies $Y(X^j)$ grows to Y_0=0.2*N for j=1 .. Y_0.
In the ideal case cumulative gain will be then a linear curve reaching Y_0 and then remaining constant; lift will then be the ratio $Y(X^j)$/0.2*j.

Cumulative gains chart

There is no ideal model, so check your predictions $P(\omega_+|X;M)$ against reality. If the prediction was true plot next point one unit to the right, one unit up; if prediction was false plot it just one unit to the right. The vertical axis contains P_+ portion of all data samples Y_0 = the number of all that responded (in this case Y_0=1000) out of all N=5000 people contacted.



Rank	$P(\omega_+\|X)$	True
1	0.98	+
2	0.96	+
3	0.96	-
4	0.94	+
5	0.92	-
.....		
1000	0.08	-

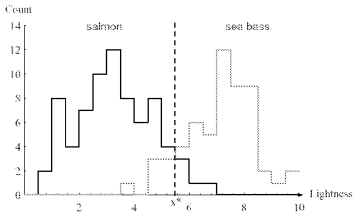
ROC intro

Receiver Operator Characteristic evaluate TP (or P_{++} rate) as a function of some threshold; for example using the likelihood ratio:

$$P(\mathbf{X}\,|\,\omega_1)\,P(\omega_1) > P(\mathbf{X}\,|\,\omega_2)\,P(\omega_2)$$

$$\Lambda(\mathbf{X}) = \frac{P(\mathbf{X}\,|\,\omega_1)}{P(\mathbf{X}\,|\,\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} = \theta$$

the threshold may be treated as a variable measuring confidence; for 1D distributions it is clear that for a large threshold only positive cases will be left, but their proportion S_+ (recall, sensitivity) decreases to zero.



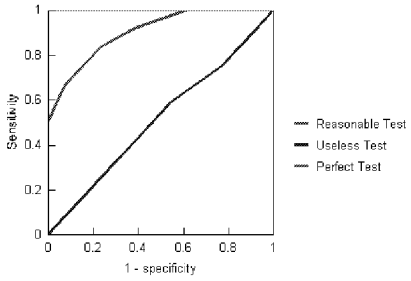
What is the optimal choice? Depends on the ratio of false alarms (FP, false positives, P_{-+}) we are willing to accept, or proportion of 1-S_-= P_{-+}/P_-

ROC curves

ROC curves display (S_+,1-S_), or error of class ω_- versus accuracy of class ω_+ for different thresholds:

Ideal classifier: below some threshold S_+ = 1 (all positive cases recognized) for 1-S_-= 0 (no false alarms).

Useless classifier (blue): same number of true positives as false alarms for any threshold.



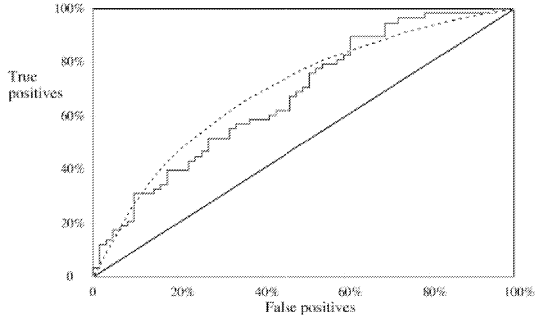
Reasonable classifier (red):

no errors until some threshold that allows for recognition of 0.5 positive cases, no errors if 1-S_- > 0.6; slowly rising errors in between.

Good measure of quality: high AUC, Area Under ROC Curve.

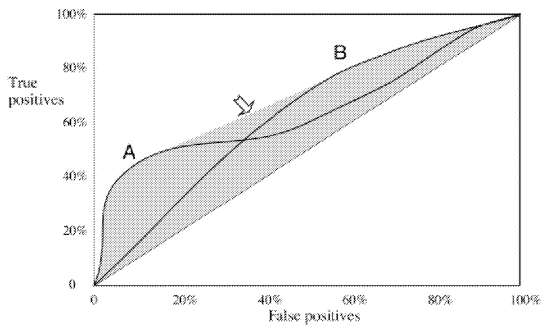
AUC = 0.5 is random guessing, AUC = 1 is perfect prediction.

Realistic ROC



ROC curve for realistic case: finite number of thresholds and data points makes it rugged. Instead of the percentage of true positives (recall, sensitivity) precision is sometimes used.
Each ROC point captures all information contained in confusion matrix for some parameters (thresholds) of the model and shows for different probabilities confidence in predictions of the classifier.
Try Yale tutorial 14 showing ROC plots.

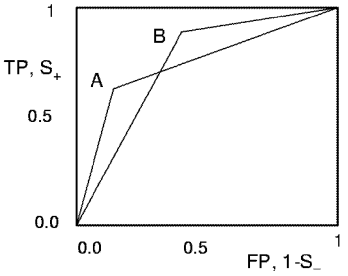
ROC for combination of classifiers

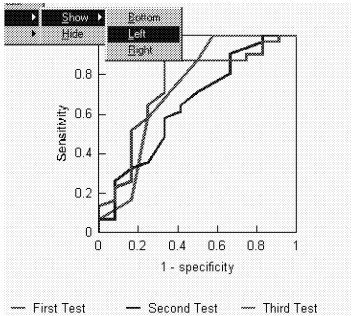


More convex ROC curves show superiority of models for all thresholds. Here ROC curves for two models show strong and weak areas: combination of the results of the two models may give a ROC curve covering the grey area.

ROC for logical rules

Rules for classifiers that give only yes/no predictions, for example logical rules, give only one point on the ROC curves, at (S_+, $1-S_-$). AUC = ($S_+ + S_-$)/2, and it is identical along $S_+ + (1-S_-) = \text{const}$ line.





Try the demo of AccuROC for Windows (sample output above)
http://www.accumetric.com/accurocw.htm