Evaluation measures

From statistics

A **confusion matrix** contains information about actual classification and predicted classification done by a classification system. Performance of such system is commonly evaluated using the data in the matrix. In general, nxn matrix can be built where an entry (i, j) contains the number of instances that belong to C_i but are assigned to C_j. Ideally, all off-diagonals should be 0, for no missclasification. The class confusion matrix allows to pinpoint what types of errors occur. The following table shows the confusion matrix for a two class classifier.

		predicted	
		yes	**no**
true classification	**yes**	TP	FN
	no	FP	TN

where

- TP (true positives) is the number of correct predictions that an instance is positive,
- FN (false negatives) is the number of incorrect predictions that an instance is negative,
- FP (false positives) is the number of incorrect predictions that an instance is positive,
- TN (true negatives)is the number of correct predictions that an instance is negative.

- The **accuracy** (AC) is the proportion of the total number of predictions that were correct:

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

- The **recall** (R) is the proportion of positive cases that were correctly identified:

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

- The **precision** (P) is the proportion of the predicted positive cases that were correct:

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

- The **F-measure** (F) combines precision and recall into a single measure of overall performance:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

where α is a factor which determines the weighting of precision and recall. A value of $\alpha = 0.5$ is often chosen. Then $F = \frac{2PR}{R+P}$

Comments

- The F-measure prefers results with more true positives, whereas accuracy is sensitive only to the number of errors. Intuitively, we are interested in finding things, even at the cost of also returning some junk.
- Using precision and recall, one can give a different cost to missing target items versus selecting junk.
- Some applications need high precision at low recall (e.g. typical web search).
- There is an obvious trade-off between precision and recall. If the whole collection is retrieved (in search document application), then recall is 100% but precision is low.

Section 1. ROC curve

ROC curve - a performance measure as a *curve* and not as a *single number statistic*.

The ROC curve (a Receiver Operating Characteristic curve) is an alternative to accuracy for the evaluation of learning algorithms. It is a plot with the false positive rate $FPR = \frac{FP}{TN+FP}$ on the x axis and the true positive rate $TPR = Recall = \frac{TP}{TP+FN}$ on the y axis.

- TPR: how many correct positives results occur among all positive examples
- FPR: how many incorrect negative results occur among all negative examples

A discrete classifier, e.g. decision trees produces a single point in ROC spase.

Many classifiers, such as Naive Bayes classifier assign to each instance $\mathbf{x}_i$ a score $f(\mathbf{x}_i)$ expressing the degree to which $\mathbf{x}_i$ is thought to be positive. In particular, Naive Bayes outputs posterior probability distributions over classes. In classification, it is often more convenient to work with scores as they can be manipulated without the need for re-normalisation.

A probabilistic classifier can be turned into a categorical classifier by setting a threshold on the score, i.e. instance $\mathbf{x}_i$ is classified as positive if $f(\mathbf{x}_i)$ is greater than a fixed threshold t, and negative otherwise. I.e.

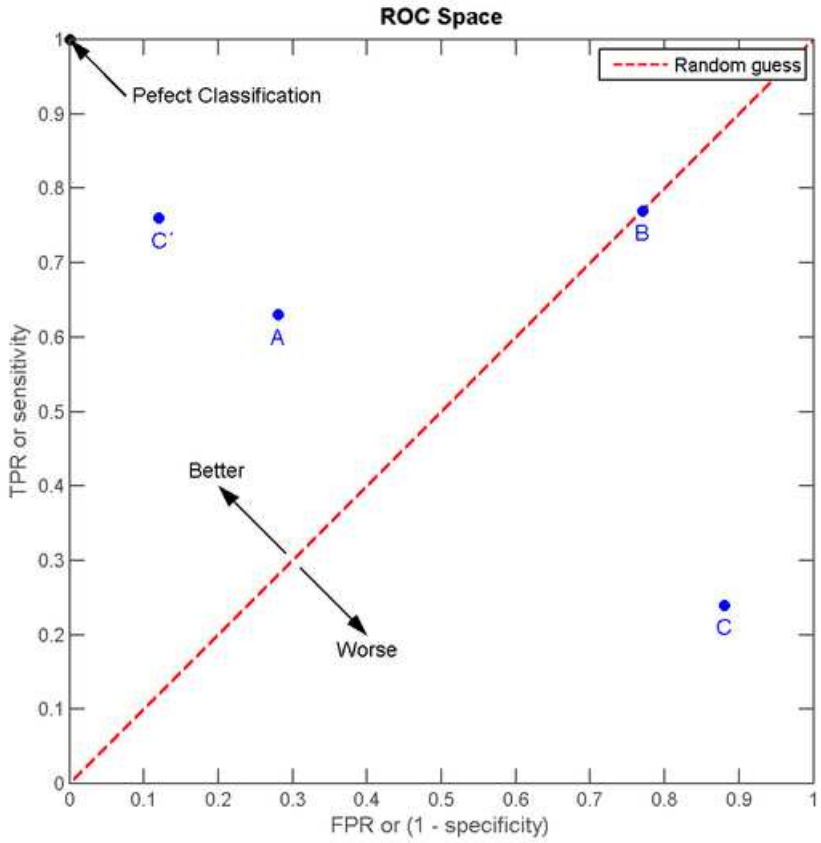
Declare $\mathbf{x}$ to be a positive if $p(y = 1|\mathbf{x}) > t$, otherwise declare it to be negative.

I.e. $y' = 1 \leftrightarrow p(y = 1|\mathbf{x}) > t$.

Number of TPs and FPs depends on threshold t. As we change t, we get differenet FPR, TPR points where $TPR = p(y' = 1|y = 1)$ and $FPR = p(y' = 1|y = 0)$.

Each parameter setting provides a (FP, TP) pair and a series of such pairs can be used to plot an ROC curve.

Figure 1



ROC

Credits: Wikipedia

A completely random guess would give a point along a diagonal line from the left bottom to the top right corners. Any point (p, p) can be obtained by predicting positive with probability p and negative with probability $1-p$.

The diagonal divides the ROC space. The upper left triangle contains classifiers that perform better than random, while the lower right triangle contains those performing worse than random.

The point **(0,1)** is the perfect classifier: it classifies all positive cases and negative cases correctly. The false positive rate is 0 (none), and the true positive rate is 1 (all).

The point **(0,0)** represents a classifier that predicts all cases to be negative, while the point **(1,1)** corresponds to a classifier that predicts every case to be positive.

The point **(1,0)** is the classifier that is incorrect for all classifications.

Classifiers mapped onto a ROC graph can be ranked according to their distance to the 'perfect performance' point. In Figure we would consider classifier C to be superior to a hypothetical classifier A because C is closer to the top left corner.

Features of ROC curves

- An ROC graph encapsulates all information contained in the confusion matrix, since FN is the complement of TP and TN is the complement of FP.
- ROC curves provide a visual tool for examining the tradeoff between the ability of a classifier to

correctly identify positive cases and the number of negative cases that are incorrectly classified.

Section 2. Tests for statistical significance

Motivation example

Assume **collocations** (btw: *Have you already started working on the final project?*): expressions consisting of two or more words that correspond to some conventional way of saying things. Like *string tea, make up, weapons of mass destructions, ...*

The simplest method for finding the collocations in a text is counting. If two words occur together a lot, then that is some evidence ... Just counting the bigrams ($w_{i-1}w_i$) does not work: *of the*, *at the*, *with the*, ...

Improving: pass the candidate phrases through part-of-speech filter which only lets through those patterns that are likely to be 'phrases'. It works pretty well!

And what about *knock* and *doors*?

- She *knocked* on his new *door*. (offset (distance) = 3 words)
- They *knocked* at the *door* .(offset (distance) = 2 words)
- 100 women *knocked* on Donaldson's *door*. (offset (distance) = 2 words)
- A man *knocked* on the metal front *door*. (offset (distance) = 4 words)

Let's employ the mean and variance based method. The mean and deviation characterize the distribution of distances between two words in a text. We can use this information to discover collocations by looking for pairs with low deviation:

- A low deviation means that the two words usually occur at about the same distance.
- Zero deviation means that the two words always occur at exactly the same distance.
- High deviation indicates that the two words of the pair stand in interesting relationship.

But ... if the two constituent words of a frequent bigram like *new companies* are frequently occuring words (as *new* and *companies* are) then we expect the two words to co-occur a lot just by chance, even if they do not form a collocation.

Let's assess whether or not something is a chance event. See **statistical testing**.

Recall the expected error rate estimation ... Now, instead of explicitly estimating some parameters, we may want to use the sample to test some particular hypothesis concerning the parameters. For example, instead of estimating mean, we may want to test if mean is less than 1.

Distinguish a notion of hypothesis in ML and hypothesis in statistical significance!

Statistical hypothesis: A statement about the parameters of one or more populations.

From statistics

Hypothesis testing

We assume a random sample of size n and we decide either to accept or to reject a statistical hypothesis on a random variable X.

A procedure for deciding to either ACCEPT or EJECT hypothesis:

- Identify the parameter of interest.
- State a null hypothesis H_0 (for ex. $H_0 : \mu = 5$).
- Specify an alternate hypothesis H_1, accepted if the null hypothesis is rejected. For ex. $H_1 : \mu \neq 5$, i.e. $H_1 : \mu < 5$ or $H_1 : \mu > 5$.
-

	True state	
	H_0	H_1
accept H_0	right decision $\mathcal{P} = 1 - \alpha$	type II error $\mathcal{P} = \beta$
reject H_0	type I error $\mathcal{P} = \alpha$	right decision $\mathcal{P} = 1 - \beta$

  - significance level $\alpha : \mathbf{Pr}(rejecting\_H_0 | H_0) = \mathbf{Pr}(type\_I\_error | H_0) = \alpha$
  - confidence level $1 - \alpha$
- Choose a significance level α. Usually $\alpha = 1$.
- State an appropriate *test statistic*: a particular statistic calculated from measurements of a random sample/experiment. A test statistic is assumed to follow a particular distribution (Normal, t, chi-square, etc.). That particular distribution can be used to test for the significance of the calculated test statistic.
- Compute the probability p (a p-value) the event related to a random variable X would occur if H_0 were true.
- If p is too low, i.e. $p < \alpha$, then reject H_0, otherwise retain H_0 as possible.

Section 3. Test statistics

Small sample test for the populatin mean: t-test

Random variables $R_1, R_2, ..., R_n$, $\overline{R} = \frac{R_1 + R_2 + ... + R_n}{n}$, distribution of R is a Normal distribution $N(\mu, \sigma^2)$ with unknown σ. Select a random sample of size n: $r_1, r_2, ..., r_n$, $\overline{r} = \frac{r_1 + r_2 + ... + r_n}{n}$.

1. H_0: $\mu = \mu_0$
2. H_1: $\mu > \mu_0$
3. Test statistics:

$$T = \frac{\overline{R} - \mu_0}{\frac{se}{\sqrt{n}}}$$

   If H_0 is valid then T has a Student's t- distribution with $n - 1$ degrees of freedom.

   t-test looks at the difference between the observed $\overline{r}$ and expected μ. It tells how likely one is to get a sample of that mean and variance assuming that the sample is drawn from a normal distribution with mean μ.

4. $t = \frac{\overline{r} - \mu}{\frac{s^2}{\sqrt{n}}} \rightarrow$ t-distribution table or calculate it yourself (see the lab session) $\rightarrow p - value$.
5. Choose a significance level α and compare α and p-value.

Example

$r_1 = 8, r_2 = 8, r_3 = 9, r_4 = 10, r_5 = 12, r_6 = 16, \overline{r} = 10.5$, $se = \sqrt{9.5}$.

1. H_0: $\mu = 9$

2. H_1: $\mu \neq 9$
3. $t = \frac{10.5-9}{\frac{se}{6}} = 1.19$
4. $\alpha = 0.05 \rightarrow t_{\frac{\alpha}{2},df=5} = 2.571$
5. $\mu \in (-2.571, 2.571)$, also $t \in (-2.571, 2.571)$.
6. H_0 is accepted with $\alpha = 5\%$.

Cross-validation

Unfortunately, training data is never large enough. So we should do our best with the data we have, i.e. mostly the small data. Let's repeat use of the **same data split** differently; this is called *cross-validation*

Let's have an data X of n instances and generate K training-test pairs $(D_i, T_i), i = 1, ..., K$.

The data X is divided randomly into K equal-sized parts $X_i, i = 1, ..., K$. To generate a training-test pair we keep one of the K parts out as the test set and combine remaining $K - 1$ parts to form the training set. Doing this K-times, we get K pairs:

$D_1 = X_2 \bigcup X_3 \bigcup ... \bigcup X_K$	$T_1 = X_1$
$D_2 = X_1 \bigcup X_3 \bigcup ... \bigcup X_K$	$T_2 = X_2$
...	...
$D_K = X_1 \bigcup X_2 \bigcup ... \bigcup X_{K-1}$	$T_K = X_K$

Leave-one-out if $K = n$.

Paired t-test

applied to a task of comparing learning algorithms.

- Two learning algorithms, L_A and L_B.
- They produce output o_A and o_B. Define random variables A and B of binomial distribution representing r proportion of misclassified instances.
- Is L_B better than L_A?

- Collect data in pairs according to the cross-validation algorithm. Repeat it K times.

instance	proportion of misclassified instances by L_A	proportion of misclassified instances by L_B
1	p_1^A	p_1^B
2	p_2^A	p_2^B
...	...	...
n	p_K^A	p_K^B

- Assume (1) The variables A, B are independent - the data is selected independently. and (2) The variables follow a normal distribution $N(\mu_A, \sigma_A)$, B follows $N(\mu_B, \sigma_B)$. The t-statistic t_α will follow a t-distribution with n-1 degrees of freedom.

Let's define a random variable T: $A - B$, $\mu_T = \mu_A - \mu_B, t_i = p_i^A - p_i^B$. We assume that differences t_i were drawn independently from a Normal distribution. We define a random variable $\overline{T} = \frac{T}{n}$ following a Normal distribution as well, $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$, s_t^2 is variance of the sample data sets.

Null hypothesis H_0: $\mu_{\overline{T}} = 0$.

We already know what we need: we can apply Student's t-test by computing

$$t_{K-1} = \frac{\bar{t} - \mu_T}{\sqrt{\frac{s_t^2}{K}}}$$

.

Thus we accept the hypothesis that two classifiers have the same error rate with $1 - \alpha$ degree of confidence if the value t_{K-1} is in the interval $(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})$.