EVALUATING LEARNING METHODS

Some Questions

How would you decide whether a machine learning program is any good?

How would you decide how good?

How would you decide if one program was better than another?

The Simple Answer

Run the program on a data set and see how well it does?

More Questions

What data set?

- Should we expect a program to perform just as well on other data sets?
- Maybe we should use several.

What do we mean by doing well?

- How do we measure performance?
- How do we decide if a performance is good?

Is it enough to run the program just once?

- If not, how often should we run it?

MEASURING PERFORMANCE

Assuming we are evaluating a system that learns to predict classifications, what would be a good measure of its performance?

Accuracy

The percentage of unknown examples that the system classifies correctly.

Not quite as simple as it seems.

In some situations, some errors matter more than others.

e.g. A system to diagnose a serious disease: What is the relative cost of a false alarm and a missed diagnosis?

Other Performance Characteristics

Accuracy is an obviously important characteristic, but there are others including:

Space and time complexity

The computing resources required by the learning procedure.

Size of training set

How many training examples are needed to reach the best achievable classification accuracy? How well will the system perform with a limited number of training examples?

Simplicity of the final model

Important if human user wants to discover something about the data.

MEASURING CLASSIFICATION ACCURACY

Suppose we have a learning program and a data set.

We use the data set to train the system.

After training the system can classify 98% of the data set correctly.

Is this good?

It is impossible to say for several reasons.

Training and Testing Must Use Different Data Sets

The system was trained and evaluated using the same data.

This does not give a good indication of how well the system will perform given a new example.

Consider an instance based system that saves all training examples.

Such a system would score 100% if tested using the data used for training!

Similar considerations apply to any learning procedure.

Correct procedure:

Randomly partition the data set into two subsets:

A *training set* used to train the system.

A *test set* used to evaluate the system's performance after learning is completed.

So let us assume we did this and the test phase achieved 95% correct classifications.

Is this good?

We Do Not Know How Difficult The Learning Task Was?

Suppose that the task was a binary classification and that 94% of examples belonged to Class 1.

> Then a system that predicted Class 1 for every unknown example would be right 94% of the time.
>
> In such a case, achieving 95% accuracy looks unimpressive.

If, on the other hand, both classes were equally likely, 95% would look very good.

So we really need to know the ***frequency of the modal class*** (i.e. the most frequently occurring class) to provide a baseline for judging the performance.

Now suppose the task is predicting the result of a coin toss on the basis of the date, time, weather and name of person tossing the coin.

> Does the fact that a learning system only achieves 50% accuracy mean the learning system is no good?
>
> Of course not. The task is such that it is impossible to predict the outcome from the attributes available.

So the ***inherent difficulty*** of a learning task must also be considered.

Generally the inherent difficulty is unknown. Hence the only available basis is *comparison with other learning procedures*.

How Many Examples Do We Need?

For training?

As many as possible.

The minimum requirement will depend on:

The complexity of the relationship that is being learned.

This will depend on both the number of attributes and the nature of the relationships between them.

More attributes will require more training data

More complex relationships between attributes will require more training data.

The learning procedure

Procedures that build more complex models require more training data.

For testing?

It depends on how accurately you want to estimate the performance

STRATIFICATION

Suppose we are training a system to predict which of two classes, C_1 and C_2, examples belong to.

Suppose also that the amount of data available is limited to a sample (which we assume is drawn randomly from the original population).

We divide a sample up into a training set and a test set.

Suppose it turns out that most of the examples in the training set belong to C_1 and most of those in the test set to C_2.

Clearly this is not a good basis for either training or testing.

What we should have done was ensure that the proportion of each class in the training set was the same as the proportion in the original sample.

This is called *stratification*.

Stratification is well worth doing but it is still possible that the training and test sets were not truly representative of the sample (and hence of the original population).

CROSS-VALIDATION

One way to further reduce the likelihood of unrepresentative training and test data sets is to repeat the process.

Suppose we divided the sample randomly into ten disjoint subsets called folds.

- Each fold could be used as the test set and the remaining 9/10 of the data set used as a training set.
- Thus we can obtain ten runs, each with a different test set, from one data set.
- The average across all 10 runs would be our accuracy estimates.

This technique is known as ***cross-validation*** and is widely used in machine learning.

Normally stratification is also imposed when partitioning the sample.

This is known as ***stratified cross-validation***.

Weka provides a facility for automatically performing n-fold stratified cross-validation. 10 way cross-validation is the default experimental procedure.

Why 10?

- Although there is no rigorous theory to back this up, wide experience suggests 10 is about the right number to get the best accuracy estimate.

EVEN BETTER ESTIMATES

A single 10-fold cross-validation will provide an accuracy estimate based on a single 10 way partitioning of the sample.

The results that you would get with a different 10 way partitioning might be different.

So, you could run the 10-fold cross validation procedure a number of times and average the results.

Note that, if you do this in Weka, you must change the random number seed manually for each cross-validation.

So how many times should you run the cross-validation procedure?

It depends how precisely you want to know the accuracy of the learning system

The Impact of Sample Size on Confidence Limits

If we assume that the results are normally distributed we can make use of the fact the standard deviation of their mean will decrease with the number of results we obtain:

$$\sigma(\bar{x}) = \frac{\sigma(x)}{\sqrt{N}}$$

This is known as the standard error.

Using this relationship we can establish a **confidence interval.**

That is, a range within which the true value will lie with some specified probability.

The width of the confidence interval *decreases with the square root of sample size*.

So we need a large number of samples if we want to estimate a parameter accurately.

Note that the required sample size is *not dependent on the population size*.

(For more details on setting up confidence intervals, using the normal distribution or Student's t distribution, see an introductory statistics text.)

CONFUSION MATRICES

So far we have considered how often a classifier gets the right answer.

However, we are sometimes also interested in what kind of mistakes it makes and how often it makes them.

Consider a diagnostic system that simply predicts whether or not a patient has a particular disease:

		Predicted	
		Yes	No
Actual	Yes	27	3
	No	17	53

This classifier is correct 80% of the time. However, the table also reveals that 17% of the predictions are false alarms (false positives) but only 3% are misses (false negatives).

		Predicted	
		Yes	No
Actual	Yes	True Positives	False Negatives
	No	False Positives	True Negatives

This type of table can be generalised to cover situations where more than two classes are predicted.

It is then known as a *confusion matrix*:

		Predicted		
		Red	Green	Blue
Actual	Red	37	1	2
	Green	3	16	11
	Blue	1	12	17

Overall, this classifier is right 70% of the time.

However, although the classifier is good at identifying red items, it is much less good at distinguishing blue and green items.

Of the 30 incorrect predictions, 23 arise from confusing blue and green.

When the system is presented with a red item, it is right 92.5% of the time.

But when the system is presented with a blue or green item, it is only right 55% of the time.

Thus the overall accuracy does not tell the whole story.

The Kappa Statistic

The classifier on the last slide predicted Red 41 times, Green 29 times and Blue 30 times:

		Predicted			
		Red	Green	Blue	Total
Actual	Red	37	1	2	40
	Green	3	16	11	30
	Blue	1	12	17	30
	Total	41	29	30	100

Suppose those predictions had been random guesses.

How often would the classifier have been right?

$$0.4 \times 41 + 0.3 \times 29 + 0.3 \times 30 = 34.1$$

So the actual success rate of 70 represents an improvement of about 35.9% on random guessing.

This is the basis of the ***kappa statistic***.

The kappa statistic expresses this improvement as *a proportion of that to be expected from a perfect predictor*.

Thus the improvement for a prefect predictor would be

100 – 34.1 = 65.9

The improvement achieved by the classifier was

70 – 34.1 = 35.9

Hence the kappa statistic is

35.9/65.9 = 0.54

A kappa statistic of 1 implies a perfect predictor.

A kappa statistic of 0 implies the classifier provides no information – it behaves as if it were guessing randomly.

Weka provides a confusion matrix and the kappa statistic in the results produced for all of its classifiers.

Recall and Precision

Recall and precision are measures originally developed in the field of information retrieval.

Consider a document retrieval system that is asked to search a set of documents for those relevant to a particular topic.

Suppose it returns a subset of documents, of which some are in fact relevant but the remainder are irrelevant.

We need some measures of how good the system is.

The system is essentially classifying the entire set of documents into two classes: relevant and irrelevant

> The *relevant* documents *returned* are examples of *True Positives* (see earlier).
>
> The *irrelevant* documents that were *returned* are examples of *False Postives*.

However we also need to consider the documents that were not returned.

> There will be *relevant* documents that were *not returned*: these are *False Negatives*.
>
> Finally, there will be *irrelevant* documents that were *not returned*: the *True Negatives*.

Information retrieval researchers have found two measures to be particularly useful in assessing the quality of an information retrieval system.

Recall

This is the most obvious measure: the proportion of relevant documents that were returned: Recall is defined as

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

where

TP is the number of relevant documents returned.

FP is the number of irrelevant documents returned.

TN is the number of irrelevant documents not returned.

FN is the number of relevant documents not returned.

Clearly, $TP+FN$ is the total number of relevant documents.

Although a high value for Recall is very desirable, it is easily achieved by a very poor system: one that returns all the documents.

Precision

For this reason we need a second measure; the proportion of the returned documents that were relevant. Precision is defined:

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

Clearly, $TP+FP$ is the number of documents returned.

Trade Off Between Recall and Precision

It is easy to build a system with 100% Recall:

> Simply return everything.

Such a system would have a very low Precision because of the large number of irrelevant documents returned.

It is almost as easy to build a system with 100% Precision.

> Only return 1 document for which the evidence of relevance is extremely strong.

Such a system would have a very low Recall because of the large number of relevant documents not returned.

So a practical information retrieval system is going to require striking a balance between Recall and Precision.

> It is almost always possible to improve one at the expense of the other.

The performance of such a system must thus be represented by a pair of numbers: a point in a 2D space.

Chi-Square Tests

Chi-square testing is an extremely useful technique for determining whether the differences between two distributions could be due to chance.

That is, whether they could both be samples of the same parent population.

Suppose we have a set of n categories and a set of observations $O_1 \ldots O_i \ldots O_n$ of the frequency that each category occurs in a sample.

Suppose we wish to know if this set of observations could be a sample drawn from some population whose frequencies we also know.

We can calculate the expected frequencies $E_1 \ldots E_i \ldots E_n$ of each category if the sample exactly followed the distribution the population.

Now compute the value of the chi-square statistic defined:

$$\chi^2 \equiv \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Clearly χ^2 increases as the two distributions deviate.

To determine whether the deviation is statistically significant, consult chi square tables for the appropriate number of degrees of freedom – in this case n-1.

MEASURES OF ASSOCIATION

We are not usually concerned only with how the values of a variable are distributed.

We generally want to know how the values of one variable depend on those of other variables.

NOMINAL VARIABLES

Suppose we collect some data by asking people in Colchester whether the think the monarchy should be abolished.

The results of our survey can be plotted in contingency table.

	Abolished	Retained
Men	60	40
Women	55	45

It appears that men are more likely to favour abolition

But

Is the effect statistically valid, or could it be an artefact of the modest sample size?

How strong is the effect?

We have already encountered a statistical test for answering the first question: the chi-square test

$$\chi^2 \equiv \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

In this case we have four observed values.

We must calculate the values we would expect if there was no difference due to gender.

There are 100 men and 100 women.

115/200 = 57.5% of the whole sample were abolitionists

95/200 = 42.5% of the whole sample were abolitionists

So if there was no effect of gender we would expect the following contingency table:

	Abolished	Retained
Men	57.5	42.5
Women	57.5	42.5

We use the chi-square test to see if the actual values are significantly different to those expected:

$$\chi^2 = \frac{(60 - 57.5)^2}{57.5} + \frac{(40 - 42.5)^2}{42.5} + \frac{(55 - 57.5)^2}{57.5} + \frac{(45 - 42.5)^2}{42.5}$$

$$= \frac{(2.5)^2}{57.5} + \frac{(-2.5)^2}{42.5} + \frac{(2.5)^2}{57.5} + \frac{(-2.5)^2}{42.5}$$

$$= \frac{6.25}{57.5} + \frac{6.25}{42.5} + \frac{6.25}{57.5} + \frac{6.25}{42.5} = 12.5\left(\frac{1}{57.5} + \frac{1}{42.5}\right) = 0.512$$

In this case there are only 2 degrees of freedom

Once the number of abolitionists is known for each gender then the number of retentionists can be inferred.

We consult the χ^2 table and find that 0.512 is far too small to provide evidence for a significant difference.

Even if there was no effect of gender there is a 75% chance that χ^2 will be as large as 0.58.

Suppose however the sample had been ten times larger and all the numbers had been multiplied by ten.

	Abolished	Retained
Men	600	400
Women	550	450

Then the calculation would become

$$\chi^2 = \frac{(600 - 575)^2}{575} + \frac{(400 - 425)^2}{425} + \frac{(550 - 575)^2}{575} + \frac{(450 - 425)^2}{425}$$

$$= \frac{(25)^2}{575} + \frac{(-25)^2}{425} + \frac{(25)^2}{575} + \frac{(-25)^2}{425}$$

$$= \frac{625}{575} + \frac{625}{425} + \frac{625}{575} + \frac{625}{425} = 125\left(\frac{1}{57.5} + \frac{1}{42.5}\right) = 5.12$$

Consulting the χ^2 tables shows that this value is significant at the 10% level but not at the 5% level.

What does this show us about the chi square test?

Any proportionate difference will be shown to be statistically significant if the sample size is large enough.

So the chi square test is very good at telling you if the evidence supports the claim that there is a difference.

But it tells you nothing about the strength of that effect.

This is rather disappointing since we are usually much more interested in larger effects.

Measuring Strength of Association

There are various measures that can be applied to measure the strength rather than the statistical significance of an association.

The Phi Coefficient

This is used only for 2x2 tables.

It is defined

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

so it is easy to compute if we already have χ^2.

Its values range between 0 (no association) and 1 (perfect association).

Note that for both our examples ϕ will have a value of 0.05 indicating an extremely weak association.

Cramer's V

For tables larger than 2x2, the phi coefficient may exceed 1 and hence is difficult to interpret.

A more general variant is Cramer's V

$$V = \sqrt{\frac{\chi^2}{N \times \min(r-1, c-1)}}$$

where r and c are the numbers of rows and columns in the contingency table.

Cramer's V has an upper limit of 1.

Lambda

Proportionate reduction of error.

An alternative to measures based on χ^2 is to consider how effective a variable is in reducing the error made in predicting another variable.

Consider following the data set on abolishing the monarchy, this time collected in Ipswich:

	Abolished	Retained
Men	60	40
Women	45	55

Suppose I know only that 52.5% of the population are abolitionists.

If I am asked to predict the opinion of a randomly chosen member of the population my best strategy would be to guess abolitionist.

Then I will be wrong 47.5% of the time.

Suppose however I know the percentages given in the table and am told the gender of the randomly selected person.

If it is a man, I will guess abolitionist and be wrong 40% of the time.

If it is a woman I will guess retentionist and be wrong 45% of the time.

Since men and women are equally likely, on average I will be wrong 42.5% of the time.

So the error rate has been reduced from 47.5% to 42.5%

Lambda is defined as the amount the error is reduced by using a category based prediction divided by the original error.

It is computed using:

$$\lambda = \frac{(\sum \max f_i) - \max F_d}{N - \max F_d}$$

where

N is the sample size

$\max f_i$ is the maximum frequency within a subclass of the independent (predicting)variable

$\max F_d$ is the modal frequency of the dependent (predicted) variable.

Lambda ranges from 0 to 1

1 implies the independent variable removes all error and permits perfect prediction.

0 implies that the independent variable supplies no additional information.

Limitations of lambda

Consider applying lambda to the original Colchester data.

Since both men and women are more likely to be abolitionists lambda will have a value of zero.

This would even be true for an extreme example such as:

	Abolished	Retained
Men	99	1
Women	51	49

NUMERICAL PREDICTIONS

So far we have been concerned with measuring the performance of classifiers. i.e. systems that predict nominal variables.

What about systems that make numeric predictions?

The key difference is that we are no longer concerned with whether predictions are right or wrong – the issue is how large the errors tend to be.

Hence, instead of accuracy (percentage correct) we typically use one of the following:

Mean Square Error

Root Mean Square Error

Mean Absolute Error

Correlation Coefficient

Coefficient of Determination – R^2

(See notes on Linear Regression)

REFERENCES

Mitchell, Chapter 5.

Witten & Frank. (2nd and 3rd editions only) Chapter 5

The basics of sampling theory are also covered by most reasonable introductory texts on statistical methods.

APPENDIX

DESCRIPTIVE STATISTICS

It is often very helpful to construct a concise summary of the main characteristics of a set of values.

Frequency Distributions

Nominal Variables:

Count the number of occurrences of each value.

Express result as percentage and/or plot as histogram.

Ordinal variables:

Can usually be treated in the same way as nominals.

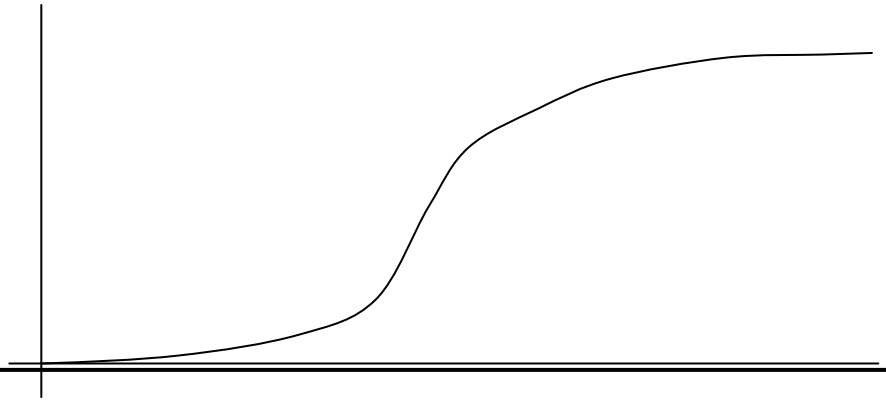
Interval-Ratio Variables:

Typically there will be a large number of distinct values.

These can be grouped into subranges and then treated in the same way as nominal variables.

Alternatively a *cumulative distribution* can be plotted:

For each value, plot the number of values less than or equal to that value.



Measures of Central Tendency

These attempt to characterise a “typical” value.

Mode.

The value which occurs *most often* in a set of values

The only possible measure for nominal variables.

Limited usefulness for other levels of variable because may not be “typical”.

Median

A value that exceeds exactly half the values in the set, but does not exceed more than half.

i.e. the “centre” of the distribution.

Not applicable to nominal values (which are unordered)

Can be used with ordinal or interval-ratio variables.

Percentiles:

Other division points can be defined:

e.g. The 90th percentile is the smallest value that exceeds 90 percent of the values.

The median is the 50th percentile

The upper quartile is the smallest value that exceeds 75% of the values.

Mean

Defined as

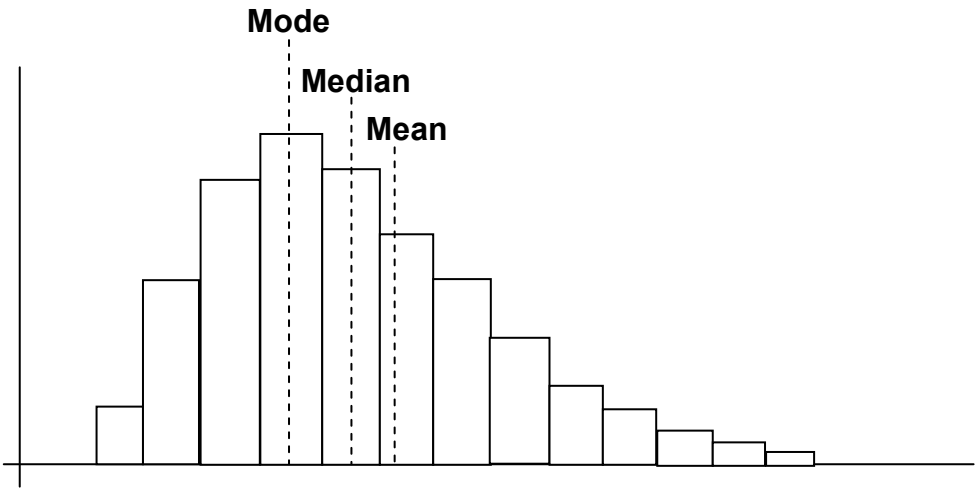
$$\overline{X} = \frac{\sum X_i}{N}$$

Can only be used with interval-ratio variables.

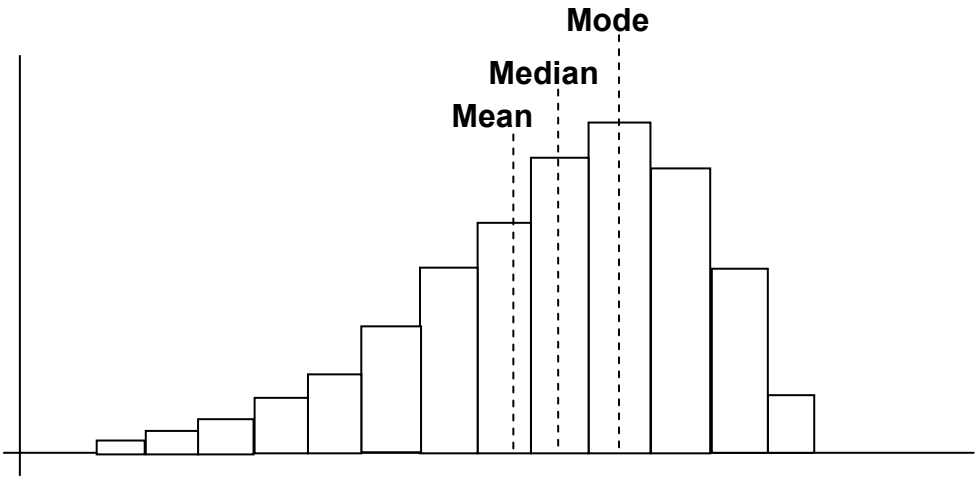
Relative positions of mode, median and mean

In a symmetrical distribution the mode, median and mean coincide (at the point of symmetry).

In an asymmetrical distribution they do not coincide:



Positively skewed distribution:



Negatively skewed distribution:

Relative merits of median and mean

The median is insensitive to changes in single values.

This property is called resistance.

It is good for representing typical values.

e.g. Median income gives a better idea of the normal standard of living than mean income.

The mean is sensitive to changes in any value.

In this sense it represents all the values in the data set.

Measures of Dispersion

These attempt to characterise the extent value to which values differ in a data set.

Index of Qualitative Variation (IQV)

Used for nominal variables

$$IQV = \frac{k(N^2 - \sum f_i^2)}{N^2(k-1)}$$

Defined as

k is the number of distinct values

f_i is the number of occurrences of the ith value

N is the total number in the data set.

IQV ranges from 0 to 1.

0 implies no variation

1 implies maximum variation

Entropy (Shannon Information)

See notes on decision trees

This measure can also be viewed as a measure of the amount of variation for a nominal variable.

Dispersion Measures for Interval Variables

Range

Defined as difference between largest and smallest values.

Depends on only two observations so very sensitive to changes in these extreme values.

Inter-Quartile Range (IQR)

Defined as difference between upper and lower quartiles

Insensitive to changes in extreme values

Like the median this is a resistant measure.

But (also like the median) does not reflect all the values

Mean Absolute Deviation (MAD)

Defined as

$$MAD = \frac{1}{N} \sum \left| X - \overline{X} \right|$$

Note that it is necessary to take the absolute values of deviations because otherwise the sum would always be zero.

Mean Squared Deviation (MSD)

As an alternative way of avoid a sum of zero, the deviations may be squared

$$MSD = \frac{1}{N} \sum (X - \overline{X})^2$$

This is mathematically more convenient than MAD.

Variance and Standard Deviation

Variance is closely related to MSD

It is defined:

$$Variance = \frac{1}{N-1}\sum (X - \overline{X})^2$$

The introduction of N-1, in place of N, reflects the fact that a single sample provides no information about the variation in the population it was taken from.

Standard Deviation

Because variance is derived from squared deviations it does not provide a direct estimate of the size of a typical deviation.

For this we use the standard deviation, s, defined:

$$s = \sqrt{Variance} = \sqrt{\frac{1}{N-1}\sum (X - \overline{X})^2}$$