Hypothesis Testing

Jayanti Prasad Ph.D

March 4, 2020

Having a model either in the form of a mathematical expression or a set of trained machine learning parameters is one of the important keys to make predictions. However, before we use a model for making predictions we must have an idea about its performance metrics. There is no single performance metric which we can use to trust a model or compare competing models. In this short note I will discuss a set of performance metrics with giving examples and explaining some use cases.

1 Introduction

Statistical modeling is used to find out the probability distribution which leads to the observational data we have as one of its realizations. Hypothesis testing is an exercise to find out whether the observational data is consistent with it. There is no way we tell whether a hypothesis is write or wrong - we can just tell that how likely it is to get the observational data point we have if a hypothesis is correct. For example, if our hypothesis is that the height of individuals in a group follows a Gaussian distribution with mean 150 cm with variance 25 cm square. If we find a person with height 250 cm then the probability of that by our hypothesis is really small.

In the present article various quanaties or 'statistics' that are used in the hypothesis testing and measuring and comparing the performance of statistical/numerical models will be dicussed.

Some of the qunatities which are discussed are accuracy, confidence measure, false aram rate, precision, recall, F1 score etc.

Hypothesis Testing

A statistical hypothesis is an assertion or conjucture about the distribution of one or more random variable [1]. In general the procedure of hypothesis testing involves the following steps:

1. Decide on the null hypothesis H_0.
2. Decide on the alternate hypothesis H_1.
3. Calculate the appropriate test statistic.
4. Decide on the significance level or the crtical P-value.
   - Type-I error : Reject H_0 when it is in fact true. The probability of this error is given by α.
   - Type-II error: Reject H_1 where in fact it is true. The probability of this errir is given by β.

   The objective in all hypothesis testing is to set the Type-I error level (significance level) at a low enough value and then use a test statistic which minimizes the Type-II error for a given sample size.
5. The P-value or critical region of size α.
6. Statement of conclusion : A decision is based on the size of the P-Value, when P-Value is too small we reject the null hypothesis.

If we have a set of observations given by a (random) variable X then the most common assumption is that it came from a Gaussian probability distribution with mean μ and variance σ^2. We can also call this our "Null Hypothesis" which will play a very important role in our discission here.

$$P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1.1}$$

Here we have an explicit expression for our Null Hypothesis but in many cases it may be just a statement like 'a person is not guilty', 'everyone who smokes have cancer' etc.

On the basis of some observed data point x_0 we want to find whether we should accept or reject our Null hypothesis given by the above probability distribution. A Gaussian probability distribution which represents our Null hypothsis is given by Figure (1.1).

For any value x we can compute the probability $P(x)$ from the above expression. If we find $P(x) < \alpha$ then we can reject the Null hypothesis. In genrtal, a small value of α is conidered for this purpose.

One of the important matrices which has been used to signify an observed data point under a given (null) hypothesis is the "p-value" which is defined in the box.

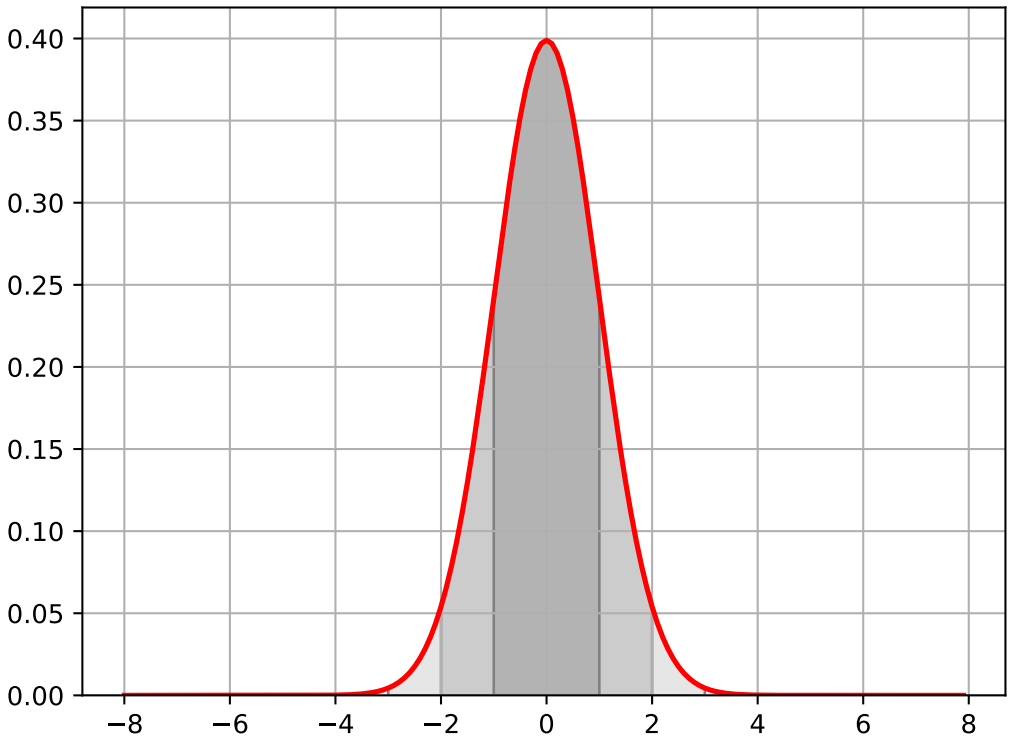

Figure 1.1: Gaussian probability distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$. The shaded regions represent the areas within 1, 2 and 2 σ which are 68.27 %, 95.45 % and 99.73 % respectively.

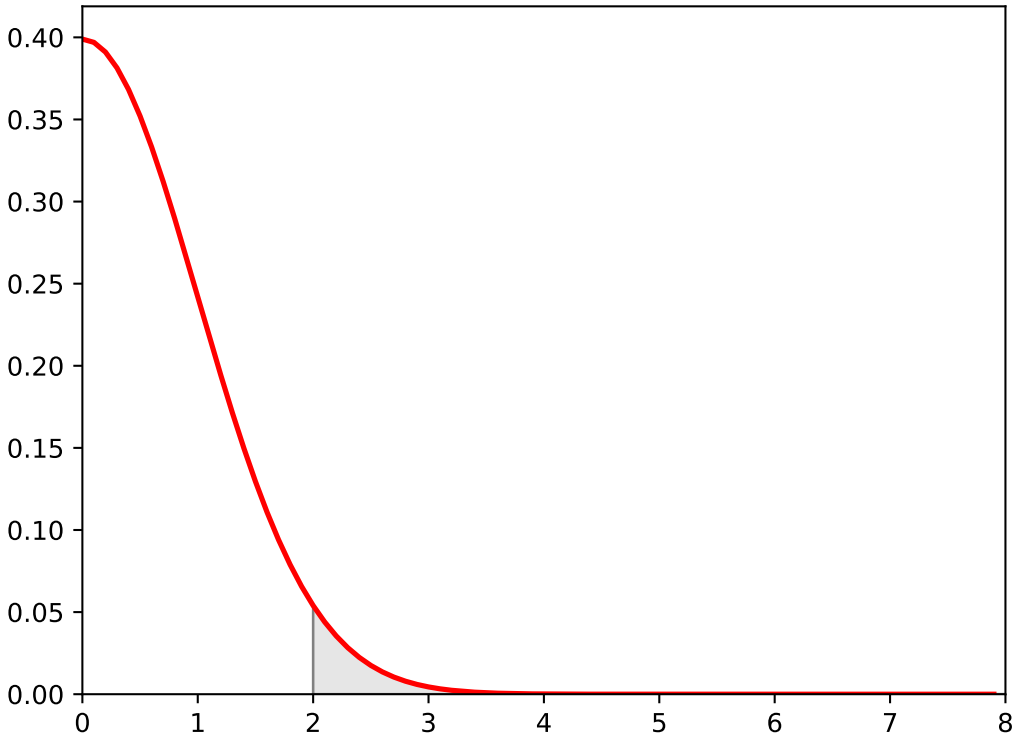

Figure 1.2: The shaded region represent the probability of obtaining a data point higher than 2σ under the Null hypothesis.

p-value

The p-value or probability value is the probability of obtaining the test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.

Note that we can have the following three cases:

- Left tail event:

$$\text{p-value} = P(X \leq x_0) \tag{1.2}$$

- Right tail event:

$$\text{p-value} = P(X \geq x_0) \tag{1.3}$$

- Double tail event:

$$\text{p-value} = 2\text{minimum}\left\{P(X \geq x_0), P(X \leq x_0; \mu, \sigma^2)\right\} \tag{1.4}$$

Looking at the probability distribution we can observe that the chances of getting a data points further away from the mean (center) becomes very small. This means that getting a p-value too small indicates that the Null hypothesis is most likely to be not true.

The null hypothesis is rejected by comparing the p-value with a user defined parameter α as mentioned above and is also called the False Alarm Rate (FAR) or False Positive (FP). The typical values which are use for α are 0.05, 0.01, 0.005, or 0.001.

The p-value is related to confidence level also. If the Null hypothesis is True then $1 - \alpha$ is the probability of making a correct decision and accepting that. This means that lower the value of α higher is our confidence. Or in the other words lesser is the False alarm rate (Type-I error) higher is our confidence.

We do not know whether a null hypothesis is True or False so we need a statistical tests. Since we may need to take decisions on the basis of whether the null hypothesis is True or False (the probability of the piece of land I own has the same probability of having gold buried as any other piece of land) and so need to carry a test.

The test can be of any type but the most common is the comparison of a measured quantity with a fixed value. It is very hard to come up with a test which will always declare a True null hypothesis as True and a False Null hypothesis as False. In most cases we will have some cases of True Null hypothesis being identified as false and vice versa. Note that a False hypothesis being identified True and a True hypothesis being identified as False are two different type of errors which we need to control as will be discussed below.

Z-tests

Z-tests can be used to compare population means to a sample's means. The z-score tells how far, in standard deviations, a data point is from the mean or average of a data set. A z-test compares a sample to a defined population and is typically used for dealing with problems relating to large samples (n $>$ 30). Z-tests can also be helpful when we want to test a hypothesis. Generally, they are most useful when the standard deviation is known.

$$z = \frac{x - \mu}{\sigma/\sqrt{n}} \tag{1.5}$$

where x is sample mean, μ population mean and σ is population standard deviation and n is the number of sample points.

t-tests

Like z-tests, t-tests are also used to test a hypothesis, but they are most useful when we need to determine if there is a statistically significant difference between two independent sample groups. In other words, a t-test asks whether a difference between the means of two groups is unlikely to have occurred because of random chance. Usually, t-tests are most appropriate when dealing with problems with a limited sample size (n $<$ 30). Both z-tests and t-tests require data with a normal distribution, which means that the sample (or population) data is distributed evenly around the mean. For a set of two samples of size n_1 and n_2 with mean x_1 and x_2, we define the t-statistic in the following way:

$$t = \frac{x_1 - x_2}{\sigma/\sqrt{n_1} + \sigma/\sqrt{n_2}} \tag{1.6}$$

A very simple example: Let's say you have a cold and you try a naturopathic remedy. Your cold lasts a couple of days. The next time you have a cold, you buy an over-the-counter pharmaceutical and the cold lasts a week. You survey your friends and they all tell you that their colds were of a shorter duration (an average of 3 days) when they took the homeopathic remedy. What you really want to know is, are these results repeatable? A t test can tell you by comparing the means of the two groups and letting you know the probability of those results happening by chance.
Another example: Student's T-tests can be used in real life to compare means. For example, a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there's always a control group (a group who are given a placebo, or "sugar pill"). The control group may show an average life expectancy of +5 years, while the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student's t-test to find out if the results are repeatable for an entire population.
The t-score is a ratio between the difference between two groups and the difference within the groups. The larger the t score, the more difference there is between groups. The smaller the t-score, the more similarity there is between groups. A t-score of 3 means that the groups are three times as different from each other as they are within each other. When you run a t-test, the bigger the t-value, the more likely it is that the results are repeatable.

- A large t-score tells you that the groups are different.
- A small t-score tells you that the groups are similar.

In the Figure (1.3) the four possibilities are shown for the hypothesis being false or true and our decision to accept or reject that. Note that we actually do not know whether the null hypothesis is true or not. Once we have p-vale we can decide whether we should accept or reject the null hypothesis by comparing that with the user decided parameter

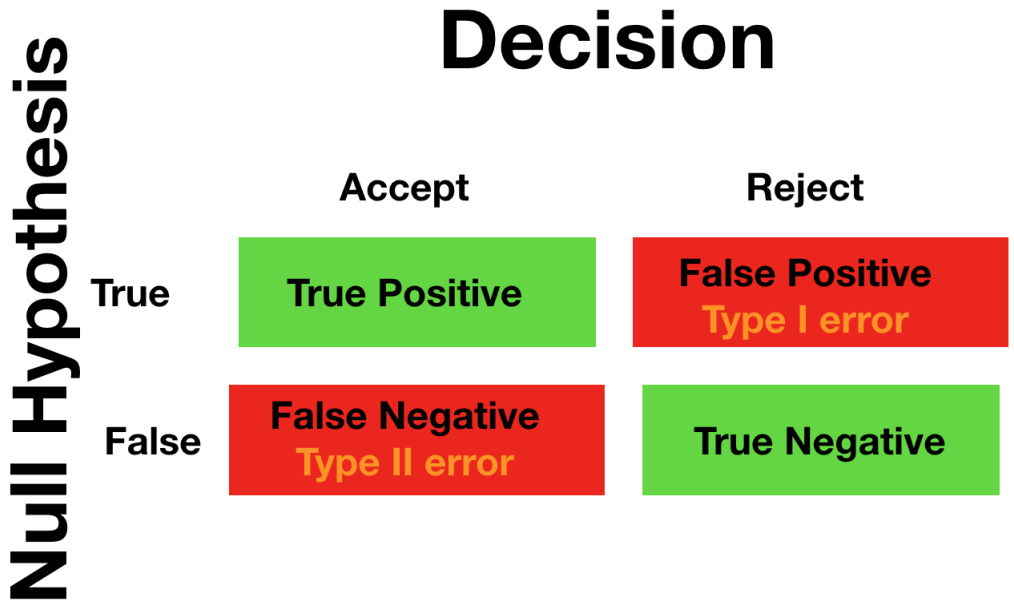

Figure 1.3: The figure shows the four possibilities between the Null hypothesis and our decision. We get Type-I error (top right red box) when we Reject a True null hypothesis and get Type-II error (bottom-left red box) when we do not reject a False Null hypothesis.

α, representing the p-value below which we reject the null hypothesis. If we make our criteria of not accepting the null hypothesis very conservative (too small value of α) we can reduce the type I error. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.

Let us consider the following situations. Someone knocks at our door and our null hypothesis is that the person is not a thief.

Type-I error

A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis. This means that your report that your findings are significant when in fact they have occurred by chance.

Type-II error

A type II error is also known as a false negative and occurs when a researcher fails to reject a null hypothesis which is really false. Here a researcher concludes there is not a significant effect, when actually there really is.

The probability of making a Type-II error is called Beta β, and this is related to the power of the statistical test (power $= 1 - \beta$). We can decrease the risk of committing a Type-II error by ensuring that our test has enough power. This can be done by using a large sample size.

- Our model predicts an innocent person as a thief and so we commit a Type-I errors.
- Our model predicts a thief an innocent person and we commit a Type-II error.

It is up to us that which error we want to keep low and that may depend on situations. For example, in a medical test if our hypothesis is "a person does not have diseses". If we make a Type-I error (declare the person ill) that is less dangerous than Type-II error (we declare an ill person healthy).

On the basis of the metrics True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) we can define the following quantities.

1. **Accuracy**
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1.7}$$
This is the most common metric.

2. **Precision:**
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.8}$$
This is also called Positive Predictive Value (PPV). If we predict that out of 8 balls predicted red only 5 are actually red then the precision will be 5/8.

3. **Recall:**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.9}$$

This is also called Sensitivity and True Positive Rate and represents compleateness of the results. Basically, this gives the fraction of the total relevent cases which have been correctly obtained.

4. **Specificity:**

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{1.10}$$

Hypothesis Testing Example

Let us assume that a factory producing food bags of 8 kg come with standard deviation of 0.16 kg. In order to check the quality if randomly pick 25 bags and find that they have mean 8.112 kg. Now we have the following hypothesis:

- Null Hypothesis:

$$\mu = 8 \tag{1.11}$$

- Alternative Hypothesis:

$$\mu \neq 8 \tag{1.12}$$

Let us consider that $\alpha = 0.01$ which corresponds 2.58 σ. Test statistic :

$$z = \frac{x - \mu}{\sigma/\sqrt{N}} = \frac{8.112 - 8.0}{0.16/\sqrt{25}} = 3.50 \tag{1.13}$$

Since $z > 2.58$ so we can reject the null hypothesis with 1 % significance level.

2 Analysis of Variance (ANOVA)

The Analysis of Variance or ANOVA can be used as an exploratory tools to explain observational data. Let us consider a data set in terms of a set or feature columns and one target columns. For example, we can have a data set of individuals with their gender and race as feature columns and their monthly income as a target variable. We do not know whether the gender has more impact on income or rance. Here we can use ANOVA to find the answer.

Here our data set is given by $D = (X_i, Y_i)$ with $i = 1, 2, .., N$ and $X_i = (G_i, R_i)$ where $G_i \in [M, F]$ and $R_i \in [A, B]$. Here the target variable Y is a continuous variable and we can compute its variance :

$$\sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{i=N} (Y_i - \bar{Y})^2 \tag{2.1}$$

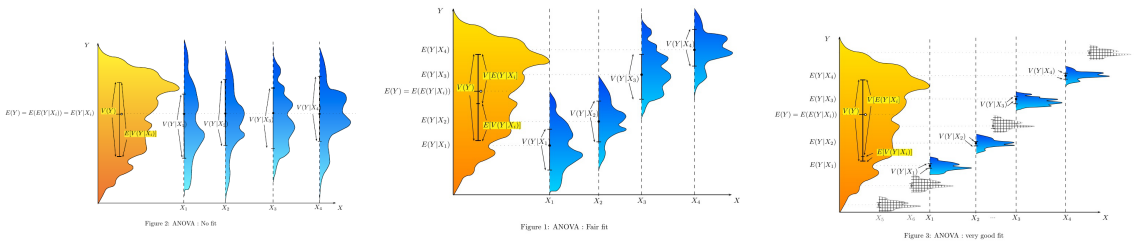

Figure 2.1: We can have well seprated mean with low variance if we use the right fetaure for grouping the data points. In the above figure the feature used in the last panel does a very good job.

Now we group our data points in the following two ways:

- Gender wise:
$$D = D_M + D_F \quad \text{and} \quad N = N_M + N_F \tag{2.2}$$
- Race Wise
$$D = D_A + D_B \quad \text{and} \quad N = N_A + N_B \tag{2.3}$$

On the basis of the above grouping we can compute μ_i and σ^2 for the groups (M, F) and (A, B).

If we find that μ for the group M and F are significantlt different and the variances σ_M^2 and σ_F^2 are significantly smaller than the other split then we can conclude that the gender is more important for the income than the race.

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) is used to analyze the differences among group means in a sample and is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal.

References

[1] Roslind L. P. Phang (1987), *Basic Concepts in Hypothesis Testing*

[2] By John A. Swets, Robyn M. Dawes, John Monahan, 2000, Scientific American October 2000, *Better Decisions through Science*

[3] Tom Fawcett (2006), Pattern Recognition Letters 27 (2006) 861–874, *An introduction to ROC analysis*

[4] Understanding ROC Curves From Scratch.