STATISTICS CONCEPTS

ANDREW MAURER

Abstract. These are just some statistical concepts that have come up several times in what I've been doing, and I figured it would be useful to have them written up for my own benefit. I might turn some of these into blog posts.

Contents

1. R^2 Score

1.1. **Definition and Intuition.** The idea behind an R^2-score (or *coefficient of determination*) is to quantify how your model compares to the most basic model possible.

Suppose you have input data X, which can be a mixture of both numerical and categorical, and they are being used to predict a numerical output Y. The simplest possible model is the average value $\bar{y}$, which is the single number which minimizes sum-squared error. Symbolically,

$$\bar{y} = \arg\min_{z} \sum_{y \in Y} (y - z)^2. \tag{1.1.1}$$

This model which identically predicts the mean value of the data-set (which I will call the *trivial model* or $\bar{y}$) is the simplest possible model, and as a base-line we would like to know how any other model compares to $\bar{y}$.

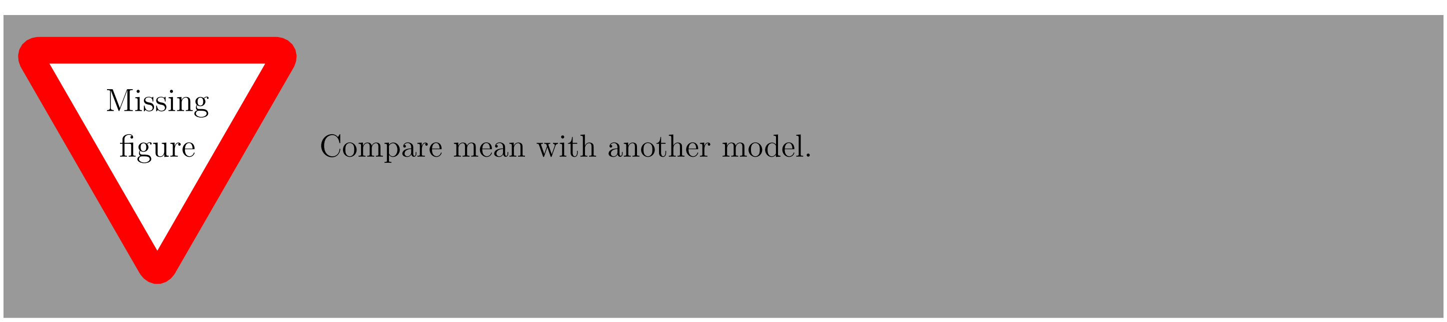


Figure 1. Two competing models

Date: July 16, 2019.

The R^2-score is how we compare a model to the trivial model. First, let us define the *sum squared error* for a model f:

$$\mathrm{SSE}(f) = \sum_{(x,y)\in D} (f(x) - y)^2 \tag{1.1.2}$$

To compare a model f to $\bar{y}$, we investigate the percentage of $\bar{y}$'s sum squared error which is still present in f's sum squared error, i.e., $\mathrm{SSE}(f)/\mathrm{SSE}(\bar{y})$. This will be a non-negative number. The R^2-score is then defined to be

$$R^2 = 1 - \frac{\mathrm{SSE}(f)}{\mathrm{SSE}(\bar{y})} \tag{1.1.3}$$

and may be interpreted as the proportion of $\bar{y}$'s error which is explained by f. Notice that $R^2 \in (-\infty, 1]$.

(1) If $R^2 = 1$, then $\mathrm{SSE}(f) = 0$ meaning we have fit the data perfectly. Be careful of overfitting.
(2) If $R^2 = 0$, then $\mathrm{SSE}(f) = \mathrm{SSE}(\bar{y})$ and all your work modeling has yielded no payoff.
(3) If $R^2 < 0$, then $SSE(f) > \mathrm{SSE}(\bar{y})$ and you need to seriously re-evaluate your life choices.

1.2. **Adjusted R^2 score.**

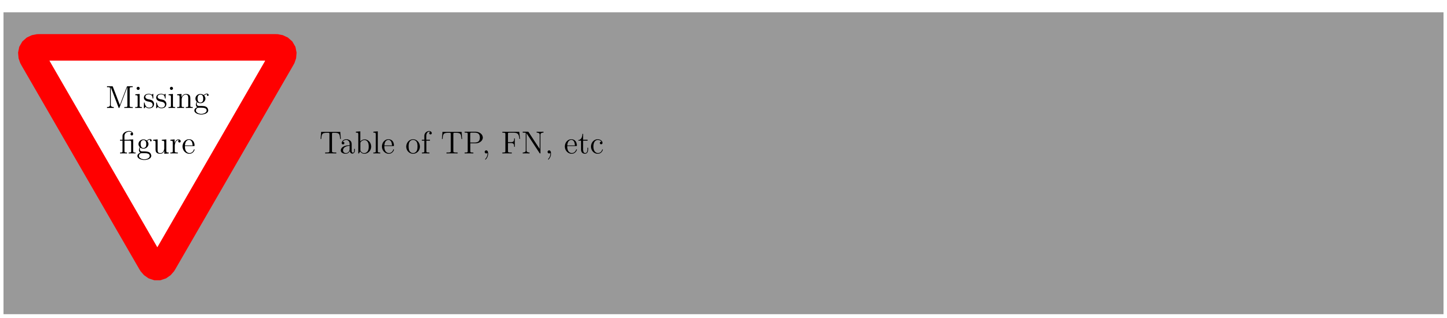
2. Receiver Operating Characteristic (ROC)

2.1. **Overview.** When performing binary classification, we often times encode binary data as 0s and 1s, and perform regression. Therefore, when predicting a value with $\hat{y} \approx 0.089$, we infer that y is most likely 1. There is some *threshold* $\alpha \in [0,1]$ for which negative predictions are those with $\hat{y} < \alpha$ and positive predictions are those with $\hat{y} \geq \alpha$.

Definition 2.1.1. In binary classification with threshold α as above, we make the following definitions for *true positive*, *false positive*, *true negative*, and *false negative* respectively.

(i) TP: predictions with $\hat{y} \geq \alpha$ and $y = 1$.
(ii) FP: predictions with $\hat{y} \geq \alpha$ and $y = 0$.
(iii) TN: predictions with $\hat{y} < \alpha$ and $y = 0$,
(iv) FN: predictions with $\hat{y} < \alpha$ and $y = 1$.

These can be arranged into a table.



With true positives et cetera defined, we may talk about the *true positive rate* and the *true negative rate*, defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \text{ and } \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2.1.1}$$

The idea behind the ROC curve is that as the threshold α changes, the TPR and TNR change. By mapping $\alpha \to (\mathrm{TPR}(\alpha), \mathrm{TNR}(\alpha))$, we parametrize a curve in the unit box $[0,1]^2 \subseteq \mathbb{R}^2$.

3. P Values

4. Hypothesis Testing

4.1. When running an experiment it is important to adhere to the scientific method in which you develop a hypothesis and test that hypothesis using data. Here I'll talk about the mathier side of

4.2. Hypotheses and Types of Errors.

4.3. Suppose a i.i.d. random variables $X_i \sim B(p)$ follow a Bernuoilli distribution with parameter $p \in [0, 1]$. The central limit theorem says that after some large number of independent trials, the mean of these trials will follow a normal distribution

$$\bar{X} = \frac{\sum X_i}{n} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Now for another set of random variables $Y_i \sim B(p')$ we would like to *test the hypothesis that $p' > p$*. To test this there are several quanitites we must specify:

- **Significance level:** The *significance level* is an upper bound on an acceptable probability of a type-1 error. In other words, the significance level is the maximum possible value of

  $$\mathbb{P}(\text{ reject } H_0 \mid H_0 \text{ is true})$$

  The significance level is often written α. The typical acceptable values of α is 0.05.
- **Power:** The *power* is a minumum ability to detect type 2 errors. In other words the power is the minimum value of

  $$\mathbb{P}(\text{ rejected } H_0 \mid H_0 \text{ is false})$$

  Closely associated is the probability of a type-2 error

  $$\beta = 1 - \text{power} = \mathbb{P}(\text{ failed to reject } H_0 \mid H_0 \text{ is false})$$

  A typical acceptable value for power is 0.80 or $\beta = 0.20$
- **Percent Improvement:** This is the level of improvement we would expect to see in the test.
- **Assignment proportions:** Some data will receive control and will become X values, some will receive treatment and become Y values. Call the proportion of individuals in the control group γ.

These three quantities allow for the computation of a number n, the number of samples necessary to detect statistically significant results.

Say n individuals will participate in this study. This means

$$\bar{X} \sim N\left(p, \sqrt{\frac{p(1-p)}{\gamma \cdot n}}\right) \text{ while } \bar{Y} \sim N\left(p', \sqrt{\frac{p'(1-p')}{(1-\gamma) \cdot n}}\right)$$

The significance level determines a cutoff C such that $\alpha = \mathbb{P}(\bar{X} > C)$, which may be unraveled into Equation 4.3.1.

$$??? \tag{4.3.1}$$

5. A/B Testing

5.1. A/B testing is a statistical method for comparing features on a website.

Email address: andrew.b.maurer@gmail.com

URL: andrewmaurer.github.io