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 $\bar{y}$.

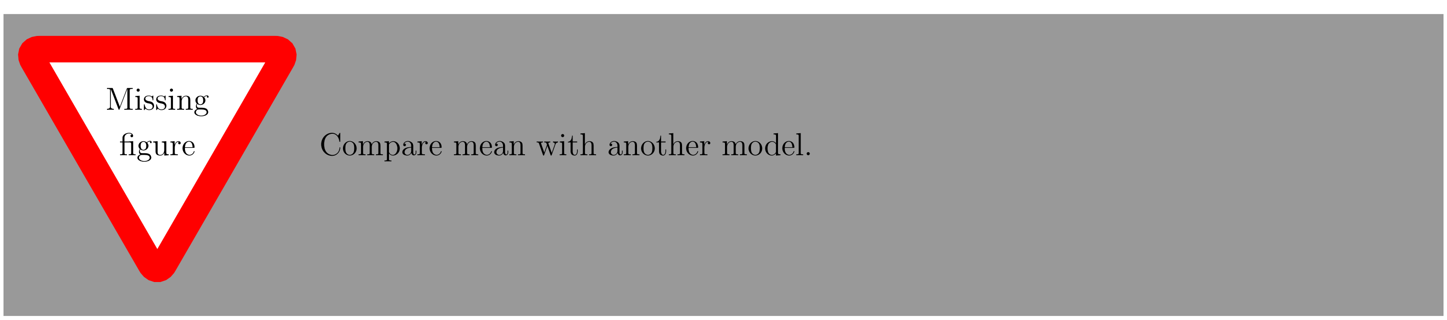

Figure 1. Two competing models

Date: March 12, 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.

Table of TP, FN, etc

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

Email address: `andrew.b.maurer@gmail.com`
URL: `andrewmaurer.github.io`