module3 · evaluation machine learning models

feature selection

Ockham's Razor states the best models are simple while fitting datasets appropriately.

A balance to achieve between accuracy and simplicity to mitigate the "curse of dimensionality"

More simple models:

- tend to predict better labels
- more interpretable to humans
- easier to make predictions from (considering less calculations)

However, selecting the optimal combinations of features is computationally difficult to apply.

Standard **Feature Selection** methods exist to remove irrelevant and redundant features from models.

Greedy Backward Selection

- Begin with all of the features in a dataset
- Find the most feature that hurts predictive power the least after removed → **remove it**
- Reiterate the process until some determined criterion is met

The process is referred to as "**greedy**" because removed features are never returned in this process.

Greedy Forward Selection

- Begin with none of the features in a dataset
- Find the single most valuable feature towards prediction power → **include it**
- Reiterate the process until some determined criterion is met

The process is effectively the converse of **Greedy Backwards Selection**. Features continually added in the process where the prediction power increase less each time, the result is **diminishing returns**. Thus the two perquisite criterion to select in the process are that of how to select features. This is typically done with the feature that **boosts accuracy** the most. The other is a **stop criterion** resulting from **diminishing returns**; typically measured with **Adjusted R^2**.

Adjusted R^2

R^2 is the measure of how well the model fits the dataset (**correlation**):

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \rightarrow \qquad \textbf{Goodness of fit}$$

If the model is always measured closely to the data, then y_i and $f(x_i)$ are close. Thus $y_i - f(x_i)$ will be ≈ 0 and R^2 will be ≈ 1 and determine a measure of **goodness of fit**. The denominator of the R^2 function does not depend on the model as it is simply a property captured of the data; it has no opinion on the model performance or correlation.

$\boldsymbol{\bar{R}^2}$ is the **Adjusted R²**. The $\boldsymbol{\bar{R}^2}$ measure penalizes $\boldsymbol{R^2}$ depending on the number of terms in a model.

$$\bar{R}^2 = R^2 - (1 - R^2)\frac{p}{n - p - 1} \quad \rightarrow \quad \textbf{Goodness of fit and complexity}$$

Therefore the **Adjusted** $\boldsymbol{R^2}$ both **sparse** and **accurate** models.

For example, if $\boldsymbol{p}$ is **large** → $\bar{R}^2 = R^2 - (1 - R^2) \times \textbf{Large Penalty} = \textbf{Small } \bar{R}^2$

However, if $\boldsymbol{p}$ is **small** → $\bar{R}^2 = R^2 - (1 - R^2) \times \textbf{Small Penalty} = ?\,\textbf{Depends on } R^2$

The reasoning influencing the **Adjusted R²** being ambiguous of a model with few features is because a model will a small amount of features might possible have a small R^2 in the first place, making the measurement of the **Adjusted R²** potentially equally small; The solutions returns to the concept of **Ockham's Razor** stating that a balance must be achieved between the correlation and the amount of features utilized in a model, in the latter context.