Model Selection: Weighing Model Complexity and Model Performance

2018-07-20

Model Complexity

In the previous activity, you ranked five classification algorithms based on their performance (prediction accuracy). In this activity, you will also consider the complexity of the models and use that, along with the performance, to select the "best" algorithm.

1. Rank order each of the five classification algorithms based on how complex the algorithms are (from simplest to most complex). Explain how you are making these rankings.

The primary goal of any classification algorithm is to make accurate classifications. A subgoal is that the classification model we adopt needs to be as simple as possible; Occam's Razor—if two models predict equally well, adopt the simpler model.

2. Select (or adopt) the best algorithm from the five candidate algorithms. To do this you will need to weigh the algorithms' classification accuracy against their complexity. Explain how you decided which algorithm to adopt.

Based on your rankings, you may have seen that more complex algorithms tend to make more accurate classifications. However, more complex algorithms do not tend to generalize as well to new datasets. This phenomenon is called *overfitting*. In order for the classification algorithm to be considered "good" it also needs to be able to accurately classify passengers from a new data set.

3. You will be given a new set of index cards (henceforth referred to as the **validation data**). For each of the five candidate algorithms, use the validation data to:

a) Fill in the classification predictions in Tables 1 and 2; and
b) Construct a classification table.

4. Rank order your models from worst to best based on the algorithms' (a) classification accuracy and (b) model complexity. (Two separate rank orderings.)

5. Which algorithm performed the "best" according to the validation dataset? Again, weigh the algorithms' classification accuracy against their complexity. Explain how you decided which algorithm to adopt.

6. Prepare a brief presentation of the results of your validation sets. In this presentation, also include a general outline of key ideas around the use of classification algorithms for making predictions.

Table 1: Classifications and Actual Fate for Passengers 1–15 in the Validation Data Set

Passenger	Model 1	Model 2	Model 3	Model 4	Model 5	Fate
Abelson, Mrs. Samuel (Hannah Wizosky)						Survived
Andersson, Mrs. Anders Johan (Alfrida Konstantia Brogren)						Died
Appleton, Mrs. Edward Dale (Charlotte Lamson)						Survived
Baclini, Miss. Marie Catherine						Survived
Berriman, Mr. William John						Died
Cacic, Mr. Luka						Died
Charters, Mr. David						Died
Coxon, Mr. Daniel						Died
Davidson, Mr. Thornton						Died
Dodge, Master. Washington						Survived
Elsbury, Mr. William James						Died
Eustis, Miss. Elizabeth Mussey						Survived
Goodwin, Master. Harold Victor						Died
Harris, Mr. George						Survived
Hart, Miss. Eva Miriam						Survived

Table 2: Classifications and Actual Fate for the Passengers 16–30 in the Validation Data Set

Passenger	Model 1	Model 2	Model 3	Model 4	Model 5	Fate
Hippach, Mrs. Louis Albert (Ida Sophia Fischer)						Survived
Homer, Mr. Harry (”Mr E Haven”)						Survived
Kallio, Mr. Nikolai Erland						Died
Kent, Mr. Edward Austin						Died
Laroche, Miss. Simonne Marie Anne Andree						Survived
Lulic, Mr. Nikola						Survived
Mudd, Mr. Thomas Charles						Died
Newell, Miss. Madeleine						Survived
Panula, Mr. Ernesti Arvid						Died
Phillips, Miss. Kate Florence (”Mrs Kate Louise Phillips Marshall”)						Survived
Sandstrom, Mrs. Hjalmar (Agnes Charlotta Bengtsson)						Survived
Stephenson, Mrs. Walter Bertram (Martha Eustis)						Survived
Torber, Mr. Ernst William						Died
Uruchurtu, Don. Manuel E						Died
Vestrom, Miss. Hulda Amanda Adolfina						Died