1 k-fache Kreuzvalidierung

Bei der k-fachen Kreuzvalidierung werden die Trainingsdaten k-mal in Testdaten und Validierungsdaten getrennt. Die Klassifizierer werden dann mit den Trainingsdaten trainiert und mit den Testdaten getestet und danach wieder angepasst. Das Testen der Klassifizierer wird mit diesen Daten vollzogen und danach wird das beste Model ausgewählt.

2 Die Lernkurve

Bei der Lernkurve werden die Korrektklassifizierungsrate des Modells als y gegenüber der Anzahl der Daten. Bei der gewünschten Klassifizierungsrate als Konstante kann es sein, dass es einen sehr großen Bias gibt, das heißt die Korrektklassifizierungsrate der Test-, sowie der Vaildierungsdaten pendeln sich weit unter der Konstante der Klassifizierungsrate ein. Bei einer sehr hohen Varianz scheinen die Kurven beider sehr langsam gegen die angepeilte Rate zu konvergieren, das heißt das Modell läuft die Gefahr einer Überanpassung. Scheinen beide Kurven schnell gegen die angepeilte Korrektklassifizierungsrate zu konvergieren hat man einen guten Kompromiss zwischen Bias und Überanpassung gefunden.

3 Abstimmung durch Rastersuche

Bei Hyperparametern der Modelle kann durch reine Brute-Force-Methoden der optimale Hyperparameter für das Modell anhand der Korrektklassifizierungsrate bestimmt werden.

4 Wahrheitsmatrix

Bei einer Wahrheitsmatrix gibt es vier Fälle. Entweder es wurde so klassifiziert, wie es sein sollte oder es wurde das Gegenteil klassifiziert. Das ergibt vier Werte, die man sehr einfach auswerten kann. Die Korrektklassifizierungsrate berechnet sich dann konkret zu

$$\mathrm{KKR} = \frac{\mathrm{PR} + \mathrm{PF}}{\mathrm{NF} + \mathrm{NR} + \mathrm{PR} + \mathrm{PF}} \tag{1}$$

Die Fehlerquote berechnet sich dann durch

$$\mathrm{FQ} = 1 - \mathrm{KKR} = \frac{\mathrm{NF} + \mathrm{NR}}{\mathrm{NF} + \mathrm{NR} + \mathrm{PR} + \mathrm{PF}} \tag{2}$$

Die Richtig-Positiv-Rate berechnet sich zu:

$$\mathrm{RPR} = \frac{\mathrm{PR}}{\mathrm{NR} + \mathrm{PR}} \tag{3}$$

Die Falsch-Positiv-Rate:

$$\mathrm{FPR} = \frac{\mathrm{PF}}{\mathrm{PF} + \mathrm{NF}} \tag{4}$$

Andere Werte zur Messung der Leistung sind die Genauigkeit

$$\mathrm{GEN} = \frac{\mathrm{PR}}{\mathrm{NR} + \mathrm{PR}} \tag{5}$$

Und die Trefferquote:

$$\mathrm{TQ} = \frac{\mathrm{PR}}{\mathrm{NF} + \mathrm{PR}} \tag{6}$$

Das $F1$,Maß berechnet sich dann zu:

$$\mathrm{F\,1} = 2\frac{\mathrm{GEN} \times \mathrm{TQ}}{\mathrm{GEN} + \mathrm{TQ}} \tag{7}$$

Für nichtbinäre Klassifizierer gelten dann die Formeln bei k Klassen:

$$\mathrm{GEN}_{mikro} = \frac{\sum_{n=1}^{k} \mathrm{PR}_n}{\sum_{n=1}^{k} (\mathrm{NR}_n + \mathrm{PR}_n)} \tag{8}$$

$$\mathrm{GEN}_{makro} = \frac{\sum_{n=1}^{k} Gen_n}{k} \tag{9}$$