Vorlesung Machine Learning und Data Mining

Übungsblatt für den 16.11.2004

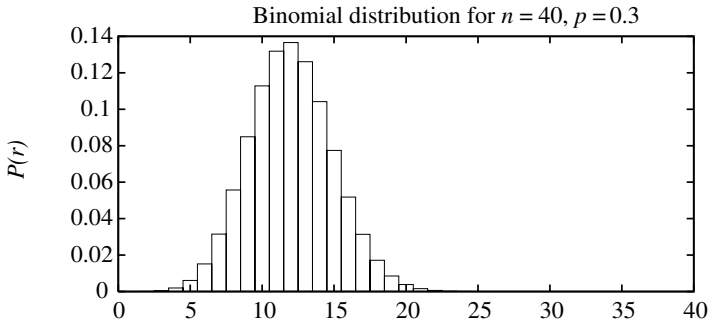
Aufgabe 1

a) Suchen Sie sich einen Gegenstand mit mind. 20 Ziffern darauf (z.B. EC/Kreditkarte, Geldscheinnummern, Studentenausweis, ...). Nehmen Sie die ersten 20 Ziffern davon und notieren Sie eine "1" für jede Ziffer, die 2 oder 5 oder 8 ist und eine "0" sonst. Zählen Sie die Anzahl der Plusse. Entspricht diese Zahl Ihrer Erwartung?

b) Erstellen Sie eine Übersicht über die Ergebnisse aller Übungsteilnehmer. Erzeugen Sie daraus ein Diagramm:

- x-Achse von 0 bis 20
- y-Achse: Tragen Sie auf, wie viele Studenten genau x Plusse hatten.

Vergleichen Sie dieses Diagramm mit dem folgenden. Worin bestehen Unterschiede?



Was würde sich an Ihrem Diagramm ändern, wenn doppelt soviele Studenten an der Übung teilgenommen hätten. Was, wenn es 10000 wären?

c) Bestimmen Sie die Punkte auf der x-Achse in Ihrem Diagramm, so daß 5%, 25%, 50%, 75% bzw. 95% des Flächeinhaltes links davon liegen. Was sagen uns diese Werte?

Aufgabe 2

a) Implementieren Sie 2 Machine-Learning-Programme x und y und denken sich 20 Datensätze mit jeweils 10000 Beispielen aus. Bestimmen Sie die Güte jedes Verfahrens auf jedem Datensatz. Welches Verfahren zur Gütebestimmung würden Sie verwenden? Wie funktioniert es? Für jeden Datensatz notieren Sie bitte eine "1" wenn die Güte von Verfahren x besser als die von y ist, sonst eine "0".

Hinweis:-))): Wenn Sie sich etwas Arbeit ersparen wollen, können Sie den Algorithmenvergleich simulieren, indem Sie die Datenreihe aus Aufgabe 1a) benutzen.

Wenden Sie nun den Vorzeichentest an um zu bestimmen, ob Ihre Verfahren x und y sich signifikant (Signifikanzniveau: 95%) unterscheiden.

b) Weicht Ihre Datenreihe aus Aufgabe 1a) signifikant (Signifikanzniveau: 95%) von der Datenreihe Ihres Nachbars ab?

Bei welchem Signifikanzniveau bemerken Sie ein Umschwenken des Ergebnisses?