Data Mining und Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Wintersemester 2015/16
Musterlösung für das 7. Übungsblatt

Aufgabe 1 Evaluierungsmethoden

Ein Datenset enthält $2 \times n$ Beispiele, wobei genau n Beispiele positiv sind und n Beispiele negativ sind. Der einfache Algorithmus `ZeroRule` betrachtet nur die Klassenverteilung der Trainings-Daten und sagt für alle Beispiele die Klasse + voraus, wenn mehr positive als negative Beispiele in den Trainings-Daten enthalten sind, und die Klasse − falls es umgekehrt ist. Bei Gleichverteilung entscheidet er sich zufällig für eine der beiden Klassen, die er dann immer vorhersagt.

- Wie groß ist die Genauigkeit dieses Klassifizierers, wenn die Verteilung der Trainings-Daten der Gesamt-Verteilung entspricht (d.h., wenn die Trainings-Daten repräsentativ sind)?
  Lösung: Geht man davon aus, dass die Verteilung der Trainings-Daten repräsentativ ist, so erreicht der Klassifizierer eine Genauigkeit von 50%, da er zufällig klassifiziert.

- Schätzen Sie die Genauigkeit von `ZeroRule` mittels Leave-One-Out Cross-Validation ab.
  Lösung: Bei Leave-One-Out CV wird ein Beispiel aus der Trainingsmenge entfernt, auf dem Rest gelernt und dann das eine Beispiel klassifiziert. Als Genauigkeit nimmt man den Mittelwert aller Beispiele. Nimmt man zB ein negatives Beispiel aus der Trainingsmenge heraus und lernt auf dem Rest, dann sagt der Klassifizierer die Klasse positiv vorher, klassifiziert also falsch (analog bei der Entnahme eines positives Beispiels). Der Klassifizierer `ZeroRule` wird also 0% Genauigkeit erreichen.

Aufgabe 2 Vorzeichentest

Sie vergleichen zwei Algorithmen A und B auf 20 Datensets und beobachten folgende Genauigkeitswerte:

Datenset	1	2	3	4	5	6	7	8	9	10
Algorithm A	0,91	0,86	0,93	0,74	0,65	0,91	0,87	0,95	0,78	0,86
Algorithm B	0,94	0,80	0,96	0,88	0,84	0,94	0,97	0,67	0,86	0,89
Datenset	11	12	13	14	15	16	17	18	19	20
Algorithm A	0,98	0,96	0,74	0,53	0,95	0,67	0,98	0,96	0,97	0,91
Algorithm B	0,87	0,90	0,79	0,51	0,96	0,69	0,79	0,98	0,98	0,76

a) Läßt sich mit Hilfe des Vorzeichentests nachweisen, ob einer der beiden Algorithmen A oder B signifikant besser ist als der andere? Nehmen sie ein vernünftiges Signifikanzniveau an.

Lösung: Als erstes zählt man die Siege und Niederlagen eines Algorithmus, z.B. von A. A gewinnt 7 mal und verliert 13 mal. Wie in der Vorlesung beschrieben (Foliensatz “Evaluation and Cost-Sensitive Learning”, Folie “Sign Test: Summary”) geht man von einer Binominalverteilung aus. Man kann nun in der Tabelle der Kritischen Häufigkeiten (Foliensatz “Evaluation and Cost-Sensitive Learning”, Folie “Table Sign Test”) nachschauen, ab welchem Wert man den Bereich unter der Kurve, der als kritisch angesehen wird, verlässt. Da die Nullhypothese von einer Gleichheit der beiden Algorithmen ausgeht, kann man sicherer sein, eine korrekte Aussage getroffen zu haben, je kleiner die Fläche unter der Kurve ist. So muss ein Algorithmus z.B. auf 30 Datenmengen mindestens auf 21 Mengen besser sein, um bei einer Irrtumswahrscheinlichkeit von 5% signifikant besser als der andere zu sein und bei einer Wahrscheinlichkeit von 1% auf 23 Mengen.

Schaut man in der Tabelle nach, so sieht man, dass B nicht signifikant besser als A ist, da er bei einer Irrtumswahrscheinlichkeit von 5% mindestens auf 15 Mengen gewinnen hätte müssen.

b) Angenommen die Nullhypothese eines Vorzeichentests zum Vergleich zweier Verfahren kann nicht zurückgewiesen werden. Bedeutet dies nun, daß die beiden Verfahren äquivalent sind bzw. daß keines der beiden besser ist?

Lösung: Da man keine Aussage über die Güte des Algorithmus treffen kann, folgt daraus weder, dass die Verfahren äquivalent sind, noch dass keines der beiden besser ist.

Aufgabe 3 Evaluierung und Kosten

Gegeben sei ein Datensatz mit 300 Beispielen, davon $2/3$ positiv und $1/3$ negativ.

a) Ist die Steigung der Isometrien für Accuracy im Coverage Space für dieses Problem < 1, $= 1$ und > 1?

Lösung: Da der Coverage nicht normiert ist, ist die Steigung von Accuracy immer $=1$ (Accuracy gewichtet positive und negative Beispiele gleich).

b) Ist die Steigung der Isometrien für Accuracy im ROC Space für dieses Problem < 1, $= 1$ und > 1?

Lösung: Da die y-Achse (tp (True Positives)-Achse) im ROC Space für die gegebene Beispielverteilung um den Faktor 2 gestaucht wird, ist die Steigung von Accuracy $= 1/2 < 1$.

c) Sie verwenden einen Entscheidungsbaum, um die Wahrscheinlichkeit für die positive Klasse zu schätzen. Sie evaluieren drei verschiedene Thresholds t (alle Beispiele mit einer geschätzten Wahrscheinlichkeit $> t$ werden als positiv, alle anderen als negativ klassifiziert) und messen folgende absolute Anzahlen von False Positives (fp) und False Negatives (fn):

t	fn	fp
0.7	40	20
0.5	30	60
0.3	10	80

Geben Sie für jeden Threshold an, für welchen Bereich des Kostenverhältnisses $\frac{c(+|-)}{c(-|+)}$ der Threshold optimal ist. Betrachten Sie auch die immer vorhandene Möglichkeit, alle (entspricht Threshold = 0.0) bzw. kein Beispiel (Threshold = 1.0) positiv vorherzusagen.

Lösung: Um die Bereiche für die verschiedenen Thresholds zu bestimmen, muss man jeden Threshold im ROC Space einzeichnen. Da im ROC Space die True Positive Rate über der False Positive Rate aufgetragen ist, muss man als erstes tp bestimmen.

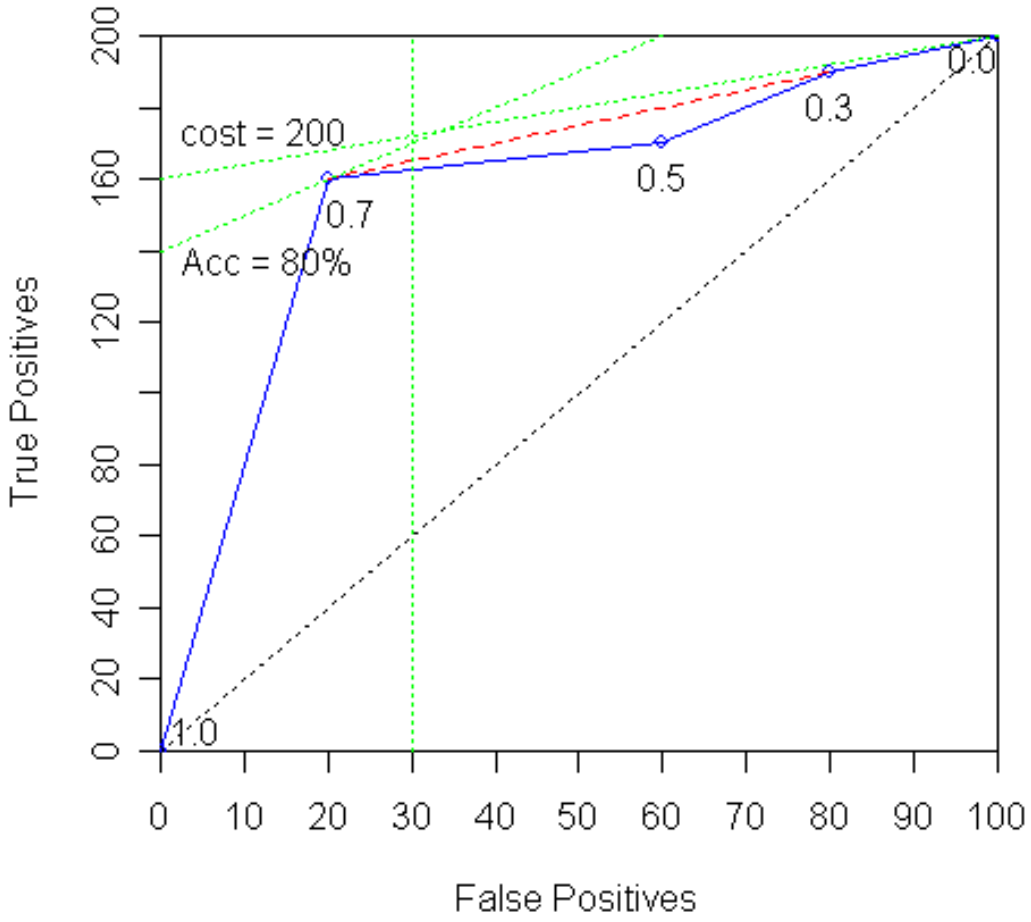
Zur Berechnung von tp rufen wir uns die Konfusionsmatrix in Erinnerung:

	classified as		
	+	-	
is +	tp (p)	fn (P-p)	P = tp+fn
is -	fp (n)	tn (N-n)	N=fp+tn

Wie man sehen kann, gilt $tp = P - fn$. Man erhält also folgende Tabelle:

t	fn	fp	tp
0.7	40	20	160
0.5	30	60	170
0.3	10	80	190

Trägt man nun die verschiedenen Thresholds in den ROC Space ein und bildet die konvexe Hülle (in der Grafik rot gestrichelt), erhält man folgende Grafik:


True Positives
False Positives
0
20
40
60
80
120
160
200
0
10
20
30
40
50
60
70
80
90
100
cost = 200
Acc = 80%
1.0
0.7
0.5
0.3
0.0

Die Länge der x und y-Achse ist gleich, wobei auf der y-Achse jeder Achsenabschnitt 20 tp's und auf der x-Achse 10 fp's entspricht. Um auf die jeweiligen "Rates" zu kommen dividiert man die Werte der x-Achse druch 100 und die der y-Achse druch 200. Um aber besser einzeichnen zu können, sind die Achsen wie in der Grafik gezeichnet.

Der Punkt (60,170), der dem Threshold 0.5 entspricht, liegt in einer Konkavität. Aus diesem Grund ist er für keinen Kosten-Bereich optimal. Man muss nun für jedes Segment der konvexen Hülle die Steigung berechnen:

- die Steigung im Abschnitt von (0,0) bis (20,160) ist $^{160}/_{20} = 8$
- die Steigung im Abschnitt von (20,160) bis (80,190) ist $\frac{190-160=30}{80-20=60} = {}^{1}/_{2}$ (da der Threshold 0.5 herausfällt)
- die Steigung im Abschnitt von (80,190) bis (100,200) ist $^{10}/_{20} = {}^{1}/_{2}$

Daraus folgt dann:

- der Threshold 1.0 ist optimal für den Bereich $\infty > \frac{c(+|-)=fp}{c(-|+)=fn} \geq 8$
- der Threshold 0.7 ist optimal für den Bereich $8 \geq \frac{c(+|-)}{c(-|+)} \geq {}^{1}/_{2}$
- der Threshold 0.3 ist nur für ein Kostenverhältnis von $^{1}/_{2}$ optimal, da die Steigung sich bis zum Punkt (100,200) nicht mehr verändert
- der Threshold 0.0 ist für Kostenverhältnisse $\leq {}^{1}/_{2}$ optimal

Zu beachten ist hier, dass sich die Steigungen auf eine nicht normalisierte Form des ROC Space beziehen, also eigentlich auf einen Coverage-Space. In diesem wären die Steigungen so wie angegeben. Da man aber das Kostenverhältnis ebenfalls in einer nicht normalisierten Form ausdrückt, ist dies ohne Auswirkung. In der Grafik kann man erkennen, dass die jeweiligen Steigungen halbiert werden müssen (da die y-Achse um den Faktor 2 gestaucht ist). Normalisiert man, muss man aber auch das Kostenverhältnis normalisieren (Bei einem Kostenverhältnis von zB $^{2}/_{5}$ hätte man in unserem Beispiel dann nach einer Normalisierung mit 2 ein Kostenverhältnis von $^{1}/_{5}$).

Visuell bzw. geometrisch kann man die Intervalle herausfinden, in dem man an dem entsprechenden Punkt des Klassifizierers eine Gerade anlegt und diese so weit um den Punkt dreht, bis der nächste eingezeichnete Punkt/Klassifizierer berührt wird, bzw. die Gerade die konvexe Hülle überlagert (zwischen dem optimalsten Punkt links oben und der Geraden dürfen sich keine weiteren Klassifizier befinden). Die größte und kleinste mögliche Steigung stellen die Intervallgrenzen dar.

d) Wie hoch ist die maximale Genauigkeit (Accuracy), die Sie im Szenario von Punkt c bei einer False Positive Rate von maximal 30% erreichen können? Wie gehen Sie dabei vor?

Lösung: Verschiebt man die Isometrien von Accuracy (also für die gegebene Beispielverteilung Linien der Steigung $1/2$) entlang der Diagonalen (($fp = 0, tp = 200$), ($fp = 100, tp = 0$)), so trifft man zuerst auf den Punkt (20,160) der dem Threshold 0.7 entspricht. Nun kann man die Accuracy für diesen Punkt errechnen (die Einschränkung, dass die $FPR < 30\%$ sein soll ist von diesem Punkt erfüllt, da hier die $FPR = 20\%$ ist).

$$Accuracy = \frac{korrekt\ klassifizierte\ Beispiele}{alle\ Beispiele} = \frac{alle\ Beispiele - (fp+fn)}{alle\ Beispiele} = \frac{300-(20+40)}{300} = 80\%$$

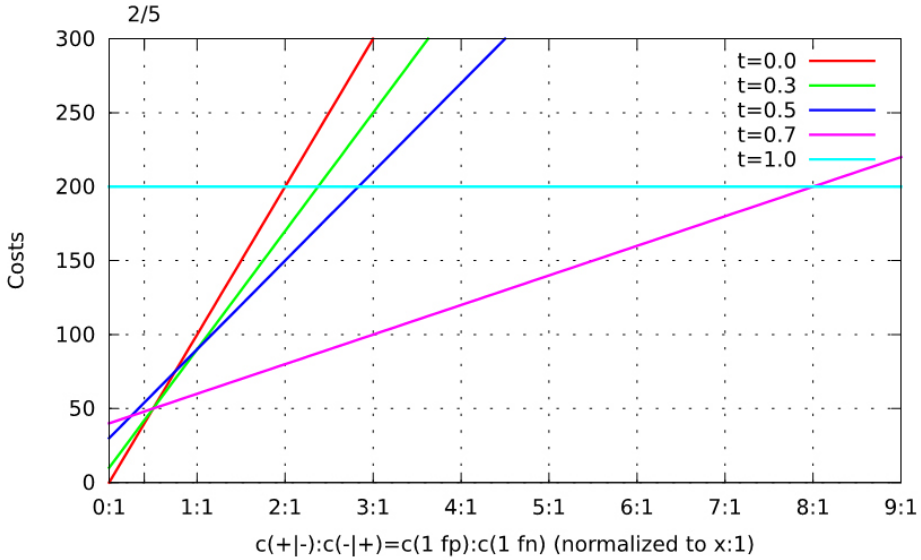
e) Sie erfahren, daß in Ihrer Anwendung ein False Positive 2 Cents kostet und ein False Negative 5 Cents kostet. Mit welchem Threshold können Sie die Kosten minimieren? Wie hoch sind die entstanden minimalen Kosten für diese 300 Beispiele?

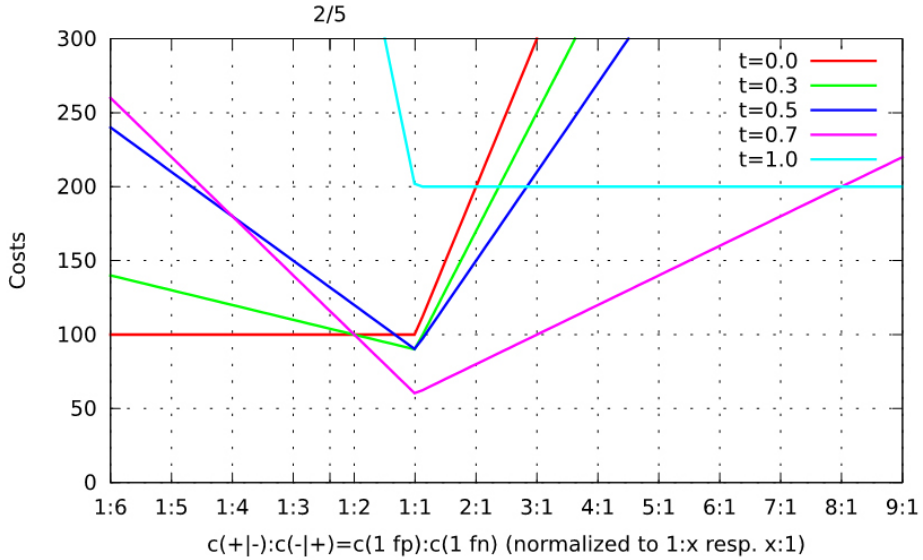
Lösung: Da man im Aufgabenteil c) berechnet hat für welche Kostenverhältnisse welcher Threshold optimal ist, kann man direkt ablesen, dass der Threshold 0.0 optimal ist, da $2/5 < 1/2$ ist. Da beim Threshold 0.0 alle Beispiele als positiv klassifiziert werden, sind alle positiven richtig klassifiziert (fn ist also 0) und alle negativen Beispiele falsch klassifiziert (fp ist also 100). Daraus folgt direkt, dass die Gesamtkosten $0 \cdot 5\ Cent + 100 \cdot 2\ Cent = 200\ Cent$ sind.

Eine andere Möglichkeit wäre, die Lösung direkt zu berechnen (also die Kosten jedes Thresholds auszurechnen und den mit den geringsten auszuwählen):

t	fn	fp	Kosten (fn c(-|+) + fp c(+|-))
1	200	0	$5 \cdot 200 + 2 \cdot 0 = 1000$
0.7	40	20	$5 \cdot 40 + 2 \cdot 20 = 240$
0.5	30	60	$5 \cdot 30 + 2 \cdot 60 = 270$
0.3	10	80	$5 \cdot 10 + 2 \cdot 80 = 210$
0.0	0	100	$5 \cdot 0 + 2 \cdot 100 = 200$

Zur besseren Veranschaulichung, im Folgenden die Kosten für die verschiedenen Thresholds in Abhängigkeit der Kostenverhältnisse. Die erste Graphik zeigt die Kosten für ein Kostenverhältnis $c_- : c_+ = c(x|-) : c(-|+)$ aufgetragen auf der x-Achse als Kostenfaktor $c_-/c_+ : 1$, die zweite Graphik zeigt zusätzlich die Kostenverhältnisse $0 < c_-/c_+ \leq 1$ als $1 : c_+/c_-$. Beachten Sie, daß man eine dreidimensionale Graphik bräuchte, um alle möglichen Kostenverhältnisse aufzeigen. Dementsprechend wird das Verhältnis 2:5 aus der Aufgabe stellvertretend durch 0.4:1 bzw. 1:2.5 dargestellt.





f) Sie bekommen die Möglichkeit, zusätzlich zu den vorhandenen 300 Beispielen noch 400 selbst auszuwählen. Wie würden Sie die Auswahl treffen, damit ein Lerner, der Kosten nicht berücksichtigen kann, unter den in Aufgabe 1e) angegebenen Kosten möglichst effektiv wird?

Lösung: Da wir nun keinen Ranker mehr haben, bei dem man einen Threshold angeben kann, sondern einen diskreten Klassifizierer, müssen wir versuchen das Kostenverhältnis über die Verteilung von positiven zu negativen Beispielen herzustellen. Da wir nun insgesamt $300+400 = 700$ Beispiele haben und das Verhältnis $2:5$ herstellen möchten, rechnen wir $2/7 \cdot 700 = 200$ und $5/7 \cdot 700 = 500$. Daher müssen wir noch $200-100 = 100$ negative und $500-200 = 300$ positive Beispiele hinzufügen, um das Kostenverhältnis von $2/5$ widerzuspiegeln.