Maschinelles Lernen: Symbolische Ansätze

Musterlösung für das 10. Übungsblatt

Aufgabe 1

Gegeben sei folgende Beispielmenge:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High		No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain		Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8		Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D11	Sunny	20	Normal	Strong	
D12	Overcast	21	High	Strong	Yes
D13		26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

a) Überlegen Sie sich eine gute Abstandsfunktion für die einzelnen Attribute.

Lösung: Wir führen für die numerischen und nominalen Attribute jeweils eine Abstandsfunktion ein. Für nominale Werte a_1 und a_2 sieht diese wie folgt aus:

$$d(a_1, a_2) = \begin{cases} 0, & \text{falls } a_1 = a_2 \\ 1, & \text{sonst} \end{cases}$$

Bei den numerischen Attribute bieten sich verschiedene Abstandsfunktionen an. Wir suchen jedoch nach einer Funktion, deren Funktionswerte mit denen der nominalen Attributen verglichen werden können. Das heißt, daß die Funktionswerte auch im Intervall $[0, 1]$ liegen sollten. Im Augenblick liegen die Werte des Attributes zwischen 8 und 29. Verwenden wir als Abstandsfunktion die Differenz

dieser Attribute, können wir diese Differenz durch das Teilen durch $21 = 29 - 8$ (bis auf Ausnahmen) auf das gewünschte Intervall abbilden:

$$d(a_1, a_2) = \frac{|a_1 - a_2|}{21}$$

Die endgültige Distanzfunktion ergibt sich dann aus der Abstandsfunktionen aller Attribute.

b) Benutzen Sie 3-NN zum Ausfüllen der fehlenden Werte.

Beziehen Sie hier die Klassifikation mit ein oder nicht? Warum?

Lösung: Wir betrachten beim Ausfüllen der fehlenden Werte eines Beispiels dessen eigene Klasse. Das heißt, wir suchen die k nächsten Nachbarn des Beispieles, die zu dessen Klasse gehören. Damit ist gewährleistet, daß keine Eigenschaften einer anderen Klasse übernommen werden.

D1: Betrachten wir nun die Beispiele, die zum Ausfüllen der Werte des Beispiels D1 benötigt werden, und berechnen deren Abstand zu D1:

Day	Outlook	Temperature	Humidity	Wind	Abstand
D2	Sunny	28	High	**Strong**	**2/21**
D6	Rain	12	Normal	Strong	2 14/21
D8	Sunny	25	High	**Weak**	**1/21**
D14	Rain	24	High	**Strong**	**1 2/21**
D15	Sunny	23	Normal	Weak	1 3/21

Die Beispiele D2, D8 und D14 sind die nächsten Beispiele. Wir erhalten zweimal *Strong* und einmal *Weak* und setzen deshalb den fehlenden Wert von D1 auf *Strong*.

D5: Diesmal müssen wir einen numerischen Werte auffüllen. Hierfür bestimmen wir wiederum die k nächsten Nachbarn und berechnen das Mittel der Attributwerte dieser.

Day	Outlook	Temperature	Humidity	Wind	Abstand
D3	Overcast	29	High	Weak	2
D4	Rain	23	High	**Weak**	**1**
D7	Overcast	8		Strong	3
D9	Sunny	18	Normal	**Weak**	**1**
D10	Rain	20	Normal	**Weak**	**0**
D12	Overcast	21	High	Strong	3
D13		26	Normal	**Weak**	**1**
D16	Sunny	21	Normal	**Weak**	**1**

Wie man sieht treten bei D5 zwei Probleme auf. Zum einen fehlt D7 auch ein Attributwert, wir treffen hier die Annahme, daß sich dieser (fehlende) Attributwert von dem von D5 unterscheidet (Abstand 1). Zum anderen können wir nicht

genau 3 nächste Nachbarn bestimmen, da 4 Beispiele den Abstand 1 haben. Wir können nun einfach den Mittelwert der Attributwerte der fünf nächsten Beispiele verwenden oder einfach 3 zufällig auswählen. Wir entscheiden uns für die erste Variante, bei der wir anschließend auf die nächste ganze Zahl abrunden:

$$\left\lfloor \frac{23 + 18 + 20 + 26 + 21}{5} \right\rfloor = \left\lfloor \frac{108}{5} \right\rfloor = 21$$

Wir füllen den fehlenden Wert also mit 21 auf.

D7, D13: Das Auffüllen der fehlende Werte erfolgt analog zu D1.

D11: Das Beispiel D11 entfernen wir komplett aus den Daten, da es uns keinen Nutzen für unsere Klassifikation bringt.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High	**Strong**	No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain	**21**	Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8	**High**	Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D12	Overcast	21	High	Strong	Yes
D13	**Rain**	26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

Damit erhalten wir den folgenden vollständigen Datensatz.

c) Welchen Klassifikationswert gibt k-NN für die folgende Instanz aus?

1. Outlook=Sunny, Temperature=23, Humidity=High, Wind=Strong

Testen Sie verschiedene k. Für welches k ändert sich die Klassifikation gegenüber $k = 1$?

Lösung: Wir berechnen zuerst die Abstände der einzelnen Trainingsbeispiele zum Klassifikationsbeispiel.

Day	Outlook	Temperature	Humidity	Wind	Klasse	Abstand
D1	Sunny	26	High	Strong	No	3/21
D2	Sunny	28	High	Strong	No	5/21
D3	Overcast	29	High	Weak	Yes	2 6/21
D4	Rain	23	High	Weak	Yes	2
D5	Rain	21	Normal	Weak	Yes	3 2/21
D6	Rain	12	Normal	Strong	No	2 11/21
D7	Overcast	8	High	Strong	Yes	1 15/21
D8	Sunny	25	High	Weak	No	1 2/21
D9	Sunny	18	Normal	Weak	Yes	2 5/21
D10	Rain	20	Normal	Weak	Yes	3 3/21
D12	Overcast	21	High	Strong	Yes	1 2/21
D13	Rain	26	Normal	Weak	Yes	3 3/21
D14	Rain	24	High	Strong	No	1 1/21
D15	Sunny	23	Normal	Weak	No	2
D16	Sunny	21	Normal	Weak	Yes	2 2/21

Wir sortieren die Beispiel aufsteigend nach ihrem Abstand zum Klassifikationsbeispiel.

Day	Outlook	Temperature	Humidity	Wind	Klasse	Abstand
D1	Sunny	26	High	Strong	No	3/21
D2	Sunny	28	High	Strong	No	5/21
D14	Rain	24	High	Strong	No	1 1/21
D8	Sunny	25	High	Weak	No	1 2/21
D12	Overcast	21	High	Strong	Yes	1 2/21
D7	Overcast	8	High	Strong	Yes	1 15/21
D4	Rain	23	High	Weak	Yes	2
D15	Sunny	23	Normal	Weak	No	2
D16	Sunny	21	Normal	Weak	Yes	2 2/21
D9	Sunny	18	Normal	Weak	Yes	2 5/21
D3	Overcast	29	High	Weak	Yes	2 6/21
D6	Rain	12	Normal	Strong	No	2 11/21
D5	Rain	21	Normal	Weak	Yes	3 2/21
D10	Rain	20	Normal	Weak	Yes	3 3/21
D13	Rain	26	Normal	Weak	Yes	3 3/21

Betrachten wir nun diese Tabelle sehen wir, daß das Beispiel für $k = 1$ (1 Beispiel negativ, 0 positiv) negativ klassifiziert wird und erst beim Einbeziehen des elftnächsten Beispiels (D3, 6 positiv, 5 negativ, $k = 11$) kommt es zu einer Veränderung der Klassifikation.

d) Berechnen Sie den Klassifikationswert obiger Instanz mittels abstandsgewichtetem NN (Shepards Methode).

Lösung: Die im Skript angebene Methode bezieht sich auf einen numerischen Klassenwert. Aus diesem Grund müssen wir uns überlegen wie wir diese auf eine nominale Klasse anwenden können. Hierfür gibt es mehrere Möglichkeiten.

Wir entscheiden uns für die folgende: Wir berechnen für beide Klassen getrennt die Summe der Kehrwerte der Abstände zwischen dem Trainingsbeispiel der jeweiligen Klasse und des Klassifikationsbeispieles. Anschließend normieren wir diese Summen, indem wir sie durch ihre Summe teilen.

Fangen wir mit der positiven Klasse an:

$$\begin{aligned} sum_+ &= \left(\frac{21}{23}\right)^2 + \left(\frac{21}{36}\right)^2 + \left(\frac{21}{42}\right)^2 + \left(\frac{21}{44}\right)^2 + \left(\frac{21}{47}\right)^2 + \left(\frac{21}{48}\right)^2 \\ &+ \left(\frac{21}{65}\right)^2 + \left(\frac{21}{66}\right)^2 + \left(\frac{21}{66}\right)^2 \approx 2,337 \end{aligned}$$

Analog für die negative Klasse:

$$\begin{aligned} sum_- &= \left(\frac{21}{3}\right)^2 + \left(\frac{21}{5}\right)^2 + \left(\frac{21}{22}\right)^2 + \left(\frac{21}{23}\right)^2 + \left(\frac{21}{42}\right)^2 + \left(\frac{21}{53}\right)^2 \\ &\approx 68,792 \end{aligned}$$

Man sieht jetzt schon, daß die negative die bessere (höhere) Bewertung bekommt. Der Form halber normieren wir diese Werte noch. Für positiv gilt:

$$\frac{sum_+}{sum_+ + sum_-} \approx 0,033$$

Für negativ gilt:

$$\frac{sum_-}{sum_+ + sum_-} \approx 0,967$$

Wie bereits zuvor erwähnt, wird das Beispiel mit Shepards Methode als negativ klassifiziert.

Aufgabe 2

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 3

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

Läßt sich mit Hilfe des Vorzeichentests nachweisen, ob einer der beiden Algorithmen A oder B signifikant besser ist als der andere? Folgt daraus, daß er nicht besser ist?

Lösung: Als erstes zählt man die Siege und Niederlagen eines Algorithmus, zB von A. A gewinnt 7 mal und verliert 13 mal. Wie in der Vorlesung beschrieben (Folie 13) geht man von einer Binominalverteilung aus. Man kann nun in der Tabelle der Kritischen Häufigkeiten (Folie 14) 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 zB 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. Da man keine Aussage über die Güte des Algorithmus treffen kann, folgt daraus auch nicht, dass er nicht besser ist.

Aufgabe 4

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

Es handelt sich bei dieser Aufgabe um die Klausuraufgabe aus dem WS 05/06. Dort ist auch eine Musterlösung dieser Aufgabe zu finden, wobei die hier vorliegende etwas ausführlicher ist.

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 x-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 und False Negatives:

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.

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.

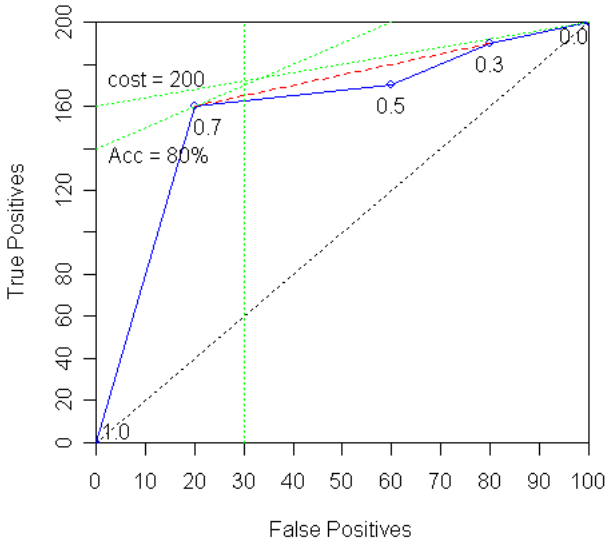
Zur Berechnung von tp rufen wir uns die Konfusionsmatrix in Erinnerung:

	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, erhält man folgende Grafik:



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 wäre die Steigung 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 $1/5$).

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 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
1	200	0	$5 \cdot 200 + 2 \cdot 0 = 1000$
0.7	40	20	$5 \cdot 40 + 2 \cdot 20 = 220$
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	$5 \cdot 0 + 2 \cdot 100 = 200$

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 Punkt e 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.