Vorlesung Wissensentdeckung

Klassifikation und Regression: nächste Nachbarn

Katharina Morik, Uwe Ligges

Informatik LS 8
Computergestützte Statistik
Technische Universität Dortmund

14.05.2013

Gliederung

Funktionsapproximation

- Die beiden vorgestellten Verfahren zu maschinellem Lernen, lineare Modelle und k-nächste Nachbarn, sind Instanzen der Funktionsapproximation.
- Gegeben sind die Trainingsbeispiele $\mathcal{T}$, gesucht ist eine Funktion

$$f_\theta(\vec{x}) = \sum_{k=1}^{K} h_k(\vec{x})\theta_k$$

.
- Dabei gibt es Parameter θ, die abzuschätzen sind, bei den linearen Modellen ist dies β.
- Darüber hinaus können die Daten transformiert werden in einen Raum, der für das Lernen besser geeignet ist: $h_k(\vec{x})$.
- Optimiert wird ein Qualitätskriterium, z.B. wird eine Verlustfunktion minimiert oder die Wahrscheinlichkeit maximiert.

Wege der Funktionsapproximation

Verlustfunktion: Fehler minimieren als Abstand zwischen wahrem Wert und Ergebnis der gelernten Funktion, z.B. $RSS(\theta)$ minimieren. Das haben wir bisher gesehen.

Likelihood: Wahrscheinlichkeit der wahren Werte maximieren! Das schauen wir uns jetzt an.

Maximum Likelihood

Gegeben eine Verteilung $Pr_\theta(y)$ und eine Stichprobe dieser Verteilung $y_1, ..., y_N$, ist die logarithmierte Wahrscheinlichkeit:

$$L(\theta) = \sum_{i=1}^{N} \log Pr_\theta(y_i) \tag{1}$$

Genau das θ, das y_i am wahrscheinlichsten macht, ist gut – $L(\theta)$ maximieren!

- Wir können dafür eine Verteilung annehmen, da wir die wahre Verteilung nicht kennen.
- Meist ist die Normalverteilung eine gute Annahme.

Normalverteilung $\mathcal{N}(\mu, \sigma)$

normalverteilt

Eine Zufallsvariable X heißt **normalverteilt** mit den Parametern μ, σ, wenn sie die Dichtefunktion

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}((x-\mu)/\sigma)^2} \tag{2}$$

besitzt.

Normalverteilung

Die zugehörige Wahrscheinlichkeitsverteilung $X \sim \mathcal{N}(\mu, \sigma^2)$ heißt **Normalverteilung**, der Graph ihrer Dichtefunktion wird Gaußsche Glockenkurve genannt.

Bei linearen Modellen ist die Maximum Likelihood gleich der Minimierung von RSS

Wir wollen θ schätzen, so dass die richtige Ausprägung von Y auch die wahrscheinlichste ist, gegeben X, θ. Unter der Annahme der Normalverteilung:

$$Pr(Y|X,\theta) = \mathcal{N}(f_\theta(X), \sigma^2)$$

Nun entspricht die log-likelihood der Daten gerade $RSS(\theta)$:

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{N} log(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i - f_\theta(\vec{x}_i)}{\sigma})^2}) \\ &= C_2 + C_1 \cdot \sum_{i=1}^{N} (y_i - f_\theta(\vec{x}_i))^2 \end{aligned}$$

Wie das?

Herleitung von $L(\theta) = RSS(\theta) \cdot C_1 + C_2$ bei Normalverteilung

$$
\begin{aligned}
L(\theta) &= \sum_{i=1}^{N} log(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y_i - f_\theta(\vec{x}_i)}{\sigma})^2}) \\
&= \sum_{i=1}^{N} \left(log(1) - log(\sigma\sqrt{2\pi}) + log(e^{-\frac{1}{2}(\frac{y_i - f_\theta(\vec{x}_i)}{\sigma})^2}) \right) \\
&= \sum_{i=1}^{N} \left(0 - log(\sigma) - log(\sqrt{2\pi}) - \frac{1}{2}(\frac{y_i - f_\theta(\vec{x}_i)}{\sigma})^2 \right) \\
&= \underbrace{-N \cdot log(\sigma) - \frac{N}{2} log(2\pi)}_{=:C_2} \underbrace{- \frac{1}{2\sigma^2}}_{=:C_1} \sum_{i=1}^{N} (y_i - f_\theta(\vec{x}_i))^2 \\
&= RSS(\theta) \cdot C_1 + C_2
\end{aligned}
$$

N, σ sind konstant für einen Datensatz.

Log-likelihood bei nominalem Y ist Entropie

Cross-Entropie

Sei Y eine Zufallsvariable, die als Werte die Namen von K verschiedenen Klassen annimmt.

$Pr(Y = y_k|X = \vec{x}) = p_{k,\theta}(\vec{x}), k = 1, ..., K$

$$L(\theta) = \sum_{i=1}^{N} log(p_{y_i,\theta}(\vec{x}_i)) \tag{3}$$

Wenn man $L(\theta)$ maximiert, passt θ gut zu den Daten im Sinne der Likelihood.

Modellselektion

- Wir haben zwei Modellklassen gesehen: lineare Modelle und Nächste Nachbarn.
- Bei der Verallgemeinerung zur Funktionsapproximation haben wir außerdem Basisfunktionen zur Vorverarbeitung gesehen, die ebenfalls Modellklassen induzieren.
- Wie wählen wir nun Modelle aus?

Verfahren zur Modellselektion

- Kreuzvalidierung für verschiedene Modelle – das mit dem geringsten durchschnittlichen Fehler nehmen! (Minimierung der Verlustfunktion jetzt auf der Ebene der Modelle)
- Direkt anhand der a posteriori Wahrscheinlichkeit Modelle vergleichen. (Maximierung der Wahrscheinlichkeit jetzt auf der Ebene der Modelle)
  - Bayes Information Criterion
  - Minimum Description Length

Kreuzvalidierung zur Modellselektion

Gegeben eine Klasse von Modellen $f(\vec{x}, \alpha)$, wobei α ein Modell der Klasse indiziert, eine Verlustfunktion $L(y, f(\vec{x}, \alpha))$, N Beispiele und eine Aufteilung der Beispiele in K Partitionen mit der Indexfunktion $\kappa : \{1, ..., N\} \to \{1, ..., K\}$, die für jede Beobachtung die zugehörige Partition angibt.

Kreuzvalidierung für alle Modelle:

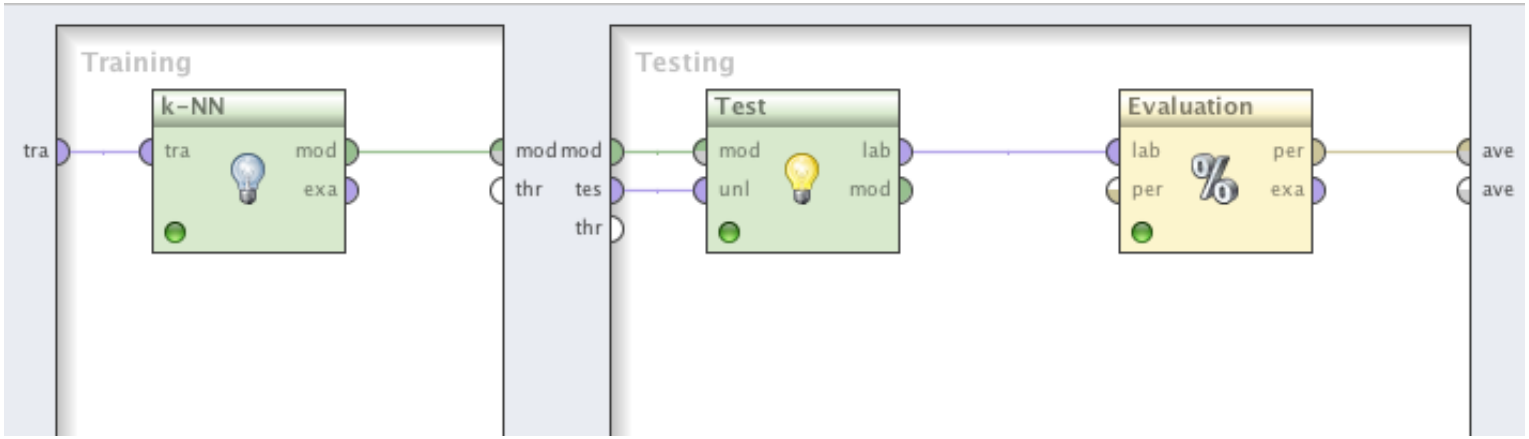
- Lasse die $\kappa(i)$-te Partition aus,
- lerne das α-te Modell: $\hat{f}^{-\kappa(i)}(\vec{x}, \alpha)$.
- rechne den Fehler aus:

$$CV(\alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(\vec{x}_i, \alpha))$$

- Minimiere $CV(\alpha)$, wähle also das Modell mit dem geringsten Verlust.

Modellselektion über Kreuzvalidierung praktisch

In RapidMiner wird die Kreuzvalidierungsschleife schon angeboten.



Es geht aber auch anders...

Bayes Statistik

A posteriori Wahrscheinlichkeit

Gegeben eine beliebige Einteilung von X in Klassen $y_1, y_2, ..., y_K$ und eine Beobachtung $\vec{x} \in X$. Die Wahrscheinlichkeit von y_j unter der Bedingung, dass $\vec{x}$ beobachtet wird, ist

$$Pr(y_j|\vec{x}) = \frac{Pr(y_j)Pr(\vec{x}|y_j)}{Pr(\vec{x})} \tag{4}$$

$Pr(y_j)$ ist die **a priori** Wahrscheinlichkeit der Klasse. $Pr(y_j|\vec{x})$ ist die **a posteriori** Wahrscheinlichkeit der Klasse.

Bayes Modellselektion

Gegeben eine Menge von Modellen $\mathcal{M}_m, m = 1, ..., M$ mit entsprechenden Parametern θ_m, Trainingsdaten $\mathcal{T}$ und eine Verteilung $Pr(\theta_m|\mathcal{M}_m)$, dann ist die a posteriori Wahrscheinlichkeit eines Modells

$$Pr(\mathcal{M}_m|\mathcal{T}) \sim Pr(\mathcal{M}_m) \cdot Pr(\mathcal{T}|\mathcal{M}_m)$$

Gegeben dass $Pr(\mathcal{M}_l|\mathcal{T}) \neq 0, Pr(\mathcal{T}|\mathcal{M}_l) \neq 0, Pr(\mathcal{M}_l) \neq 0$:

Zum Vergleich zweier Modelle $\mathcal{M}_j, \mathcal{M}_l$ berechnen wir den Quotienten:

$$\frac{Pr(\mathcal{M}_m|\mathcal{T})}{Pr(\mathcal{M}_l|\mathcal{T})} = \frac{Pr(\mathcal{M}_m)}{Pr(\mathcal{M}_l)} \cdot \frac{Pr(\mathcal{T}|\mathcal{M}_m)}{Pr(\mathcal{T}|\mathcal{M}_l)}$$

Ist das Ergebnis > 1, nehmen wir $\mathcal{M}_m$, sonst $\mathcal{M}_l$.

Approximieren der a posteriori Wahrscheinlichkeit

Wenn alle Modelle a priori gleich wahrscheinlich sind, müssen wir nur $Pr(\mathcal{T}|\mathcal{M}_i)$ approximieren.

- Mit Maximum Likelihood schätzen wir $\hat{\theta}_i$.
- Die Anzahl freier Parameter in $\mathcal{M}_i$ nennen wir d_i. Das ist z.B. die Dimension der Beispiele, kann aber wegen $h_k(\vec{x})$ oder einiger Eigenschaften des Lernverfahrens auch etwas anderes sein.
- Als Wahrscheinlichkeit nähern wir an:

$$log\ Pr(\mathcal{T}|\mathcal{M}_i) = log\ Pr(\mathcal{T}|\hat{\theta}_i, \mathcal{M}_i) - \frac{d_i}{2} \cdot log\ N + O(1) \quad (5)$$

Maximale a posteriori Wahrscheinlichkeit und BIC

Bayes Informationskriterium

Sei d die Anzahl der Parameter eines Modells und N die Anzahl der Beispiele, dann ist das Bayes Informationskriterium BIC

$$BIC = -2\ loglik + (log\ N) \cdot d \tag{6}$$

Dabei ist $loglik = \sum_{i=1}^{N} \log Pr_{\hat{\theta}}(y_i)$.

BIC als Qualitätskriterium bei Likelihood Maximierung wählt eher einfache Modelle. Unter einer Gaußschen Verteilung und bei bekannter Varianz σ^2 rechnen wir

$$-2\ loglik \sim \sum_i \frac{(y_i - \hat{y}_i)^2}{\sigma^2}$$

Die Wahl des Modells mit kleinstem BIC entspricht der Wahl des Modells mit größter a posteriori Wahrscheinlichkeit.

Relative Qualität der Modelle per BIC

- Die Wahl des Modells mit kleinstem BIC ist zuverlässig. Gegeben eine Familie von Modellen, darunter das richtige, konvergiert die Wahrscheinlichkeit, dass BIC das richtige wählt, gegen 1, wenn die Anzahl der Beispiele gegen ∞ konvergiert.
- Wenn wir für jedes Modell $\mathcal{M}_m, m = 1, ..., M$ den BIC ausrechnen, können wir (wie bei Kreuzvalidierung auch) die Modelle relativ zueinander bewerten, hier:

$$\frac{e^{-\frac{1}{2} \cdot BIC_m}}{\sum_{l=1}^{M} e^{-\frac{1}{2} \cdot BIC_l}} \tag{7}$$

Minimum Description Length

Ein Modell kodiert eine Menge von Beispielen. Wir können Nachrichten so kodieren, dass keine Nachricht Präfix einer anderen ist, z.B.

Nachricht	z1	z2	z3	z4
Code	0	10	110	111

Wir wollen den kürzesten Code für die häufigste Nachricht. Der Code des Beispiels ist optimal, wenn $Pr(z1) = 1/2$, $Pr(z2) = 1/4$, $Pr(z3) = 1/8$, $Pr(z4) = 1/8$.
Wieso das?

Shannon/Weaver Theorem

Code-Länge als Entropie

Wählen wir die Code-Länge l_i einer Nachricht z_i als

$$l_i = -log_2 Pr(z_i)$$

so ist die durchschnittliche Nachrichtenlänge

$$length \geq -\sum Pr(z_i) log_2(Pr(z_i)) \tag{8}$$

Wenn $p_i = A^{-l_i}$, wobei A die Anzahl der verwendeten Zeichen ist, gilt sogar die Gleichheit (s. Beispiel):
$Pr(z_1) = 1/2 = 2^{-1} = A^{-l_1}, A = 2, l_1 = 1$

Minimum Description Length zur Modellselektion

Gegeben ein Modell $\mathcal{M}$ mit Parametern θ und Beispiele $\mathcal{T} = (\mathbf{X}, \mathbf{y})$, der Empfänger kennt alle $\mathbf{X}$ und soll die $\mathbf{y}$ empfangen. Dazu müssen wir den Unterschied zwischen Modell und wahren Werten sowie die Modellparameter übermitteln.

Prinzip der Minimum Description Length MDL

Wähle immer das Modell mit der kürzesten Nachrichtenlänge!

$$length = -log\ Pr(\mathbf{y}|\theta, \mathcal{M}, \mathbf{X}) - log\ Pr(\theta|\mathcal{M}) \tag{9}$$

Eigenschaften von MDL

- Bei normalverteilten y, θ, wenn wir $\mathbf{X}$ zur Einfachheit weglassen, sehen wir den Einfluss von σ:

$$length = log\ \sigma + \frac{(y-\theta)^2}{\sigma^2} + \frac{\theta^2}{2}$$

- Je kleiner σ desto kürzer die Nachricht und einfacher das Modell!

Bezug zwischen MDL und BIC

- Wenn wir die Länge (Gleichung 9) minimieren

$$length = -log\ Pr(\mathbf{y}|\theta, \mathcal{M}, \mathbf{X}) - log\ Pr(\theta|\mathcal{M})$$

  maximieren wir auch die a posteriori Wahrscheinlichkeit (vgl. Gleichung 4) $Pr(\mathbf{y}|\mathbf{X})$.
- Mit Hilfe des BIC haben wir Modelle für die Funktionsapproximation durch Maximum Likelihood ausgewählt: das Modell mit dem kleinsten BIC entspricht dem Modell mit größter a posteriori Wahrscheinlichkeit.
- Also kann man das Modell mit der kleinsten Code-Länge (MDL-Prinzip) auch durch die Minimierung des BIC finden.

Was wissen Sie jetzt?

- Funktionsapproximation optimiert eine Qualitätsfunktion.
  - Fehlerminimierung, z.B. RSS, MSE
  - Maximierung der Likelihood, z.B. durch Approximation der a posteriori Wahrscheinlichkeit
    - Fehlerminimierung RSS entspricht Maximum Likelihood, falls Normalverteilung gegeben (Regression).
- Für die Modellselektion kann man
  - die Kreuzvalidierung mit Fehlerminimierung und
  - die Kriterien nach Bayes (BIC, MDL) nutzen.