Validation

Oktober, 2013

Lernziele

- Konzepte des maschinellen Lernens
  - Validierungsdaten
  - *Model Selection*
  - Kreuz-Validierung (*Cross Validation*)

Outline

Idee

Offene Frage: Welche "Modell Komplexität" ist passend, bzw. wie stark soll die Regularisierung (λ) sein?

Nicht-Trainingsdaten Daten (*out-of-sample*) zur Schätzung des *out-of-sample* Fehlers E_{out}.

$$\mathbb{E}[loss(h(\mathbf{x}), y)] = E_{out}(h)$$

Validierungsdaten

Aufteilen der m gelabelten Daten $\mathcal{D}$ zum Training mit Validierung in

- m_{train} Trainingsdaten $\mathcal{D}_{train}$
- m_{val} Validierungsdaten $\mathcal{D}_{val}$

$$\mathcal{D}_{val} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(m_{val})}, y^{(m_{val})})\}$$

Somit sind $m_{train} = m - m_{val}$ Daten für das eigentliche Training (Anpassen der Parameter) im Validierungsfall vorhanden.

Validierungsfehler

$$E_{val}(h) = \frac{1}{m_{val}} \sum_{i=1}^{m_{val}} loss(h(\mathbf{x}^{(i)}), y^{(i)}))$$

$$\mathbb{E}\left[E_{val}(h)\right] = \frac{1}{m_{val}} \sum_{i=1}^{m_{val}} \mathbb{E}\left[loss(h(\mathbf{x}^{(i)}), y^{(i)}))\right] = E_{out}(h)$$

Zuverlässigkeit des Validierungsfehler als Schätzung für E_{out}

Für die Abhängigkeit der Zuverlässigkeit des Validierungsfehler als Schätzung für E_{out} von der Anzahl der Validierungsdaten m_{val} gilt:

$$E_{val}(h) = E_{out}(h) \pm \mathcal{O}(\frac{1}{\sqrt{m_{val}}})$$

Datenverwendung: Training vs. Validierung

Verwendung der Daten $\mathcal{D}$ unter $m = m_{train} + m_{val}$

- je mehr Daten für das Training , desdo kleiner wird E_{out}; vgl. Lern-Kurven
- je mehr Daten für die Validierung, desdo zuverlässiger wird E_{val} als Schätzung für E_{out}

Daumenregel: ca. 20% der Daten für Validierung

Training/Validierung und zweites Training

1. Training und Validierung
   - Training mit m_{train}-Daten $\rightarrow h^-$
   - Validierung mit m_{val}-Daten zur Schätzung von E_{out} , d.h. Beurteilung von h^-
2. Training mit voller Datenmenge $m \rightarrow h$

$$E_{out}(h) \leq E_{out}(h^-) \leq E_{val}(h^-) + \mathcal{O}(\frac{1}{\sqrt{m_{val}}})$$

Outline

Validierung vs. Test

- Testdaten: Nur zur Abschätzung der Qualität
- Validierungsdaten werden indirekt beim Training verwendet:
  - Einstellungen der Hyperparameter
  - Verschiedene Vorverarbeitungsmethoden
  - Auswahl zwischen verschiedenen Modellen (*Modell Selection*)

Modelle

Verschiedene "Modelle" sind:

- Stärke der Regularisierung (Einstellung des Regularization Hyper-Parameter λ
- Wahl zwischen verschiedenen Lernverfahren, wie Logistische Regression, SVM, NN
- Wahl innerhalb eines Verfahrens:
  - zwischen linearem Modell, Grad- des Polynoms etc.
  - bei Neuronalen Netzen: Anzahl der Neurone in den Schichten, Anzahl der Schichten
  - bei SVM: Art des Kernels

Werden verschiedene Modelle mittels der Validierungsdaten ausgewählt, ist der Validierungsfehler E_{val} nicht mehr ein "neutraler" Schätzer des *out-of-sample errors* E_{out}.

Erläuterung am Beispiel zweier Hypothesen h_1^-, h_2^- mit $E_{out}(h_1^-) = E_{out}(h_2^-) = C$
Wahl der Hypothese h_*^- mit dem niedrigerem E_{val} führt zu $\mathbb{E}[E_{val}(h_*^-)] < C$

Mehrere Hypothesen $h_1^-, h_2^-, \ldots, h_M^-$, also $|\mathcal{H}_{val}| = M$
Auswahl der Hypothese $h_{m^*}^-$ mit dem niedrigsten E_{val}:

$$E_{out}(h_{m^*}^-) \leq E_{val}(h_{m^*}^-) + \mathcal{O}\left(\sqrt{\frac{\ln M}{m_{val}}}\right)$$

Lernen mit m Trainingsbeispielen ergibt somit (vgl. Lernkurven):

$$E_{out}(h_{m^*}) \leq E_{out}(h_{m^*}^-) \leq E_{val}(h_{m^*}) + \mathcal{O}\left(\sqrt{\frac{\ln M}{m_{val}}}\right)$$

Outline

Kreuz-Validierung

Ziel (wie bei der Validierung): Modellselektion
statt ein Modell h^- mit $\mathcal{D}_{train}$ zu trainieren und dies mit der Validierungsmenge $\mathcal{D}_{val}$ einzuschätzen:
Durchführung dieser Prozedur v-mal: v-fache Kreuz-Validierung (*v-fold cross validation*):

- Aufteilen der Daten $\mathcal{D}$ in v-Teile
- Benutzen von $v-1$ Teilen zum Training und 1 Teil zur Validierung
- v-fache Durchführung (v-mal trainieren und validieren), wobei jeweils ein anderer Teil der Daten der Validierungsdatensatz ist.
- Durchschnittsbildung des Validierungsfehlers aus den v-Validierungsfehler zur Beurteilung des Modells.

Extremfall *one-leave-out*: $v = m_{train}$, d.h. Validierung mit jeweils nur einem Datum.

Outline

Underfitting - High Bias

- Benutze ein Model mit einer höheren Kapazität (*capacity*)
- Füge aussagekräftige Features hinzu
- Erniedrige den *Regularization*-Parameter

Overfitting - High Variance

- Reduzieren der Anzahl der Features (Feature Selection)
- Benutze ein Model mit einer niedrigeren Kapazität (*capacity*)
- Erhöhe die Anzahl der Trainingsdaten
- Erhöhe den *Regularization*-Parameter

Literaturangabe

- Andrew Ng: Machine Learning (Cousera Online Kurs), 2013
- Yaser Abu-Mostafa: Learning from Data, Caltech Machine Learning bzw.
  Yaser Abu-Mostafa et all.: Learning from Data, AMLBook 2012

Weiterführende Literatur:

- Trevor Hastie,Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning, insb.: Kapitel 7, Springer Verlag 2009