Support Vector Machines

Basierend auf Blanchard, G. Bousquet, O. Massart, P. (2008)

Jens Nußberger

26. Mai 2020

Gliederung

Rückblick

Vortrag vom 12. Mai: Einführung in Statistical Learning nach Vapnik

Mustererkennung

- $\mathcal{Y} = \{0, 1\}$
- $f(\cdot, \alpha)$, $\alpha \in \Lambda$ Indikatorfunktionen
- $L(y, f(x, \alpha)) = \begin{cases} 0 & \text{falls } y = f(x, \alpha) \\ 1 & \text{falls } y \neq f(x, \alpha) \end{cases}$
- Dann gilt:

$$\begin{aligned} R(\alpha) &= \int L(y, f(x, \alpha)) \mathrm{d}\mathbb{P}(x, y) \\ &= \int \mathbb{1}\{y \neq f(x, \alpha)\} \mathrm{d}\mathbb{P}(x, y) \\ &= \mathbb{P}(y \neq f(x, \alpha)) \end{aligned}$$

- $R(\alpha)$ gibt Wahrscheinlichkeit eines Klassifikationsfehlers an.

Key Theorem of Learning Theory

- Sei $R(\alpha)$ beschränkt, d.h. es gibt $m, M \in \mathbb{R}, s.d.$

$$m \leq R(\alpha) \leq M \quad \forall \alpha \in \Lambda$$

- Dann ist das ERM Prinzip genau dann strikt konsistent, wenn $R(\alpha)$ einseitig gleichmäßig gegen $R_{emp}(\alpha)$ konvergiert.

$$\inf_{\alpha \in \Lambda(c)} R_{emp}(\alpha) \xrightarrow[n \to \infty]{\mathbb{P}} \inf_{\alpha \in \Lambda(c)} R(\alpha) \quad \forall c \in \mathbb{R}$$

$$\Leftrightarrow$$

$$\mathbb{P}\left(\sup_{\alpha \in \Lambda} (R(\alpha) - R_{emp}(\alpha)) > \varepsilon \right) \xrightarrow[n \to \infty]{} 0$$

Heute:

- Beispiel zur „Mustererkennung“: Support Vector Machine (SVM)
- Alternatives Resultat zur Fehlerabschätzung eines Empirischen Prozesses

Notation

Es werden die folgenden Begriffe verwendet:

- W-Raum: $(\Omega, \Sigma, \mathbb{P})$
- *Input* und *Output* Maßräume: $\mathcal{X}, \mathcal{Y} = \{-1, 1\}$
- Zufallsvariable $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ mit Bildmaß P
- *Training set*: $(X_i, Y_i)_{i=1}^n$ i.i.d nach P
- Empirische Verteilung: $\mathrm{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{(X_i, Y_i)}$

Außerdem:

- P_X als Randverteilung von X unter P
- $\eta(x) := \mathrm{P}(Y = 1 \mid X = x)$

für $\mathrm{P}(Y \in B \mid X = x)$, eine reguläre Version der bedingten Verteilung mit

$$\mathbb{E}[\dots] = \int f(X) \mathbb{P}(Y \in B \mid X) \, \mathrm{d}\mathbb{P}(\omega) = \int f(x) \, \mathrm{P}(Y \in B \mid X = x) \, \mathrm{d} \, \mathrm{P}_X(x) = \mathrm{E}_X[\dots],$$

für alle $\sigma(X)$-messbaren $f : \mathcal{X} \to \mathbb{R}$.
Wir identifizieren außerdem P mit (η, P_X).

Notation

...

- P_X als Randverteilung von X unter P
- $\eta(x) := \mathrm{P}(Y = 1 \mid X = x)$

für $\mathrm{P}(Y \in B \mid X = x)$, eine reguläre Version der bedingten Verteilung mit

$$\mathbb{E}[\dots] = \int f(X)\mathbb{P}(Y \in B \mid X)\,\mathrm{d}\mathbb{P}(\omega) = \int f(x)\,\mathrm{P}(Y \in B \mid X = x)\,\mathrm{d}\,\mathrm{P}_X(x) = \mathrm{E}_X[\dots],$$

für alle $\sigma(X)$-messbaren $f : \mathcal{X} \to \mathbb{R}$.
Wir identifizieren außerdem P mit (η, P_X):

Lemma

Für alle $C \subset \mathcal{X} \times \mathcal{Y}$ gibt es $C_{-1}, C_1 \subset \mathcal{X}$, so dass

$$\mathbb{P}((X, Y) \in C) = \int_{C_{-1}} (1 - \eta(x))d\,\mathrm{P}_X(x) + \int_{C_1} \eta(x)d\,\mathrm{P}_X(x).$$

Beweis: (Devroye et al., 1996, S.9)

Das Klassifikationsproblem

- Klassifikator: $f : \mathcal{X} \to \{-1, 1\}$ messbar
- Reellwertiger Klassifikator: $f : \mathcal{X} \to \mathbb{R}$ messbar (wobei $\operatorname{sgn}(f) \to \{-1, 1\}$)

Bestrafung der Misklassifikation durch *loss functions*:

- 0-1 loss: $\theta(f) := (x, y) \mapsto \mathbb{1}\{yf(x) \leq 0\}$

Klassifizierungsfehler:

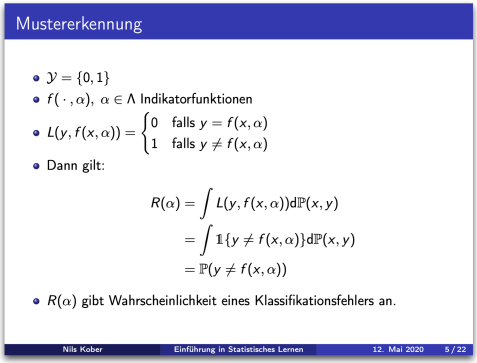
- $\mathcal{E}(f) := \mathrm{E}[\theta(f)] \overset{f \in \{-1,1\}}{=} \mathbb{P}[f(X) \neq Y]$

Optimaler Klassifikator (unter $\mathcal{E}$)

- Bayes Klassifikator: $s^*(x) = \begin{cases} 1 & \eta(x) \geq \frac{1}{2} \\ -1 & \text{sonst} \end{cases}$

Relative Risk als *relative average loss* zu s^*:

- $\Theta(f, s^*) = \mathrm{E}[\theta(f) - \theta(s^*)]$



Optimalität des Bayes Klassifikators

Lemma

Der Bayes Klassifikator $s^*(x) = \begin{cases} 1 & \eta(x) \geq \frac{1}{2} \\ -1 & \text{sonst} \end{cases}$ *ist optimal unter allen messbaren Funktionen* $s : \mathcal{X} \to \{-1, 1\}$ *in dem Sinne, dass*

$$\mathbb{P}(s^*(X) \neq Y) \leq \mathbb{P}(s(X) \neq Y).$$

Weiterhin ist s^* *fast sicher eindeutig auf der Menge* $\{\mathrm{P}(Y = 1 \mid X = x) \neq \frac{1}{2}\}$.

Beweis: (Devroye et al., 1996, S.10).

Support Vector Machines als Large Margin Classifier

Gegeben sei ein *Training set* $(X_i, Y_i)_{i=1}^n$ mit $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$.

Large Margin Classification

$$\min_{\beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}} \|\beta\|$$
$$\text{s.t. } Y_i(\langle \beta, X_i \rangle + \beta_0) \geq 1, \quad \forall i = 1, \ldots, n$$

unter Relaxation für nicht-separierbare Daten:

Einführung von *slack* Variablen ξ_i

$$\min_{\substack{\beta \in \mathbb{R}^d, b \in \mathbb{R} \\ (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n}} \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i$$
$$\text{s.t. } Y_i(\langle \beta, X_i \rangle + b) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n$$
$$\xi_i \geq 0, \quad \forall i = 1, \ldots, n$$

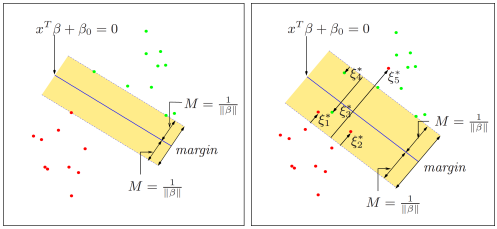


Abbildung 1: Lineare Entscheidungsgrenze mit margin für separierbare Daten (links) und nicht separierbare Daten (rechts). Quelle: Hastie et al. (2009)

Der Kernel Trick

Definition

Ein Hilbertraum $\mathcal{H}$ ist ein $\mathbb{R}$-Vektorraum mit Skalarprodukt $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, der vollständig ist bezüglich der vom Skalarprodukt induzierten Norm.
Der (Topologischer) Dualraum: $\mathcal{H}^* = L(\mathcal{H}, \mathbb{R})$ von $\mathcal{H}$, ist der Raum der beschränkten linearen Funktionale von $\mathcal{H}$ nach $\mathbb{R}$.

Definition

Eine symmetrische Funktion $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt positiv definit, wenn $\forall n \geq 1$, $\forall (a_1, \ldots, a_n) \in \mathbb{R}^n$ und $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0$$

Der Kernel Trick - RKHS

Definition

Für einen Hilbertraum $\mathcal{H} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ und ein $x \in \mathcal{X}$ heiße

$$\delta_x : \mathcal{H} \ni f \mapsto f(x)$$

das Evaluationsfunktional.

Definition

Ein Hilbertraum $\mathcal{H} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ heißt Hilbertraum mit reproduzierendem Kern (RKHS), falls $\forall x \in \mathcal{X} : \delta_x \in \mathcal{H}^*$

Definition

Eine Funktion $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ heißt reproduzierender Kern von $\mathcal{H}$ wenn

1. $\forall x \in \mathcal{X} : k(\cdot, x) \in \mathcal{H}$,
2. $\forall x \in \mathcal{X}\ \forall f \in \mathcal{H} : \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.

Insbesondere gilt $\langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} = k(x, y),\ \forall x, y \in \mathcal{X}$

Konstruktion von RKHS (Moore-Aronszajn)

Sei $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ positiv definit, dann gibt es einen eindeutigen RKHS $\mathcal{H}_k \subset \mathbb{R}^{\mathcal{X}}$ mit reproduzierendem Kern k. Weiterhin ist $\mathcal{H}_k$ der Abschluss von $\text{span}\{k(\cdot, x) \mid x \in \mathcal{X}\}$ unter dem inneren Produkt

$$\langle f, g \rangle_k = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \beta_j k(x_i, x_j),$$

wobei $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ und $g = \sum_{j=i}^{n} \beta_j k(\cdot, x_j)$.

Für mehr Theorie zu RKHS siehe Sejdinovic and Gretton (2013).

Support Vector Machines mit Kernel Trick

SVM in $\mathbb{R}^d$

$$\min_{\substack{\beta\in\mathbb{R}^d, b\in\mathbb{R} \\ (\xi_1,\ldots,\xi_n)\in\mathbb{R}^n}} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i$$

$$\text{s.t. } Y_i(\langle\beta, X_i\rangle + b) \geq 1 - \xi_i, \quad \forall i = 1,\ldots,n$$

$$\xi_i \geq 0, \quad \forall i = 1,\ldots,n$$

SVM mit Kernel Trick

$$\min_{\substack{f\in\mathcal{H}_k, b\in\mathbb{R} \\ (\xi_1,\ldots,\xi_n)\in\mathbb{R}^n}} \frac{1}{2}\|f\|_k^2 + C\sum_{i=1}^n \xi_i$$

$$\text{s.t. } Y_i(\langle f, k(\cdot, X_i)\rangle_k + b) \geq 1 - \xi_i, \quad \forall i = 1,\ldots,n$$

$$\xi_i \geq 0, \quad \forall i = 1,\ldots,n$$

Die Lösung des Optimierungsproblems ist von der Form:
$\sum_{i=1}^n Y_i\alpha_i k(x, X_i) + b = \langle f^*, \phi(x)\rangle_k + b$

Support Vector Machines als *Penalization Method*

SVM mit Kernel Trick

$$\min_{\substack{f\in\mathcal{H}_k, b\in\mathbb{R} \\ (\xi_1,\ldots,\xi_n)\in\mathbb{R}^n}} \frac{1}{2}\|f\|_k^2 + C\sum_{i=1}^n \xi_i$$

$$\text{s.t. } Y_i(\langle f, k(\cdot, X_i)\rangle_k + b) \geq 1 - \xi_i, \quad \forall i = 1,\ldots,n$$

$$\xi_i \geq 0, \quad \forall i = 1,\ldots,n$$

Schreibe SVM_0 für das SVM Problem mit $b = 0$ und wähle $\Lambda_n = \frac{1}{nC}$:

SVM_0 als *Penalization Method*

$$\min_{f\in\mathcal{H}_k} \underbrace{\frac{1}{n}\sum_{i=1}^n (1 - Y_i f(X_i))_+}_{\stackrel{\text{def}}{=} \mathrm{E}_n[\ell(f)]} + \Lambda_n\|f\|_k^2$$

Das Klassifikationsproblem

- Klassifikator: $f : \mathcal{X} \to \{-1, 1\}$ messbar
- Reellwertige Klassifikation: $f : \mathcal{X} \to \mathbb{R}$ messbar (wobei $\operatorname{sgn}(f) \to \{-1, 1\}$)

Bestrafung der Misklassifikation durch *loss functions*:

- 0-1 loss: $\theta(f) := (x, y) \mapsto \mathbb{1}\{yf(x) \leq 0\}$
- hinge loss: $\ell(f) := (x, y) \mapsto (1 - yf(x))_+$

Klassifizierungsfehler:

- $\mathcal{E}(f) := \mathrm{E}[\theta(f)] \overset{f \in \{-1,1\}}{=} \mathbb{P}[f(X) \neq Y]$

Optimaler Klassifikator (unter $\mathcal{E}$)

- Bayes Klassifikator: $s^*(x) = \begin{cases} 1 & \eta(x) \geq \frac{1}{2} \\ -1 & \text{sonst} \end{cases}$

Relative Risk als *relative average loss* zu s^*:

- $\Theta(f, s^*) = \mathrm{E}[\theta(f) - \theta(s^*)]$
- $L(f, s^*) := \mathrm{E}[\ell(f) - \ell(s^*)]$

Mustererkennung

- $\mathcal{Y} = \{0, 1\}$
- $f(\,\cdot\,, \alpha)$, $\alpha \in \Lambda$ Indikatorfunktionen
- $L(y, f(x, \alpha)) = \begin{cases} 0 & \text{falls } y = f(x, \alpha) \\ 1 & \text{falls } y \neq f(x, \alpha) \end{cases}$
- Dann gilt:

$$\begin{aligned} R(\alpha) &= \int L(y, f(x, \alpha)) \mathrm{d}\mathbb{P}(x, y) \\ &= \int \mathbb{1}\{y \neq f(x, \alpha)\} \mathrm{d}\mathbb{P}(x, y) \\ &= \mathbb{P}(y \neq f(x, \alpha)) \end{aligned}$$

- $R(\alpha)$ gibt Wahrscheinlichkeit eines Klassifikationsfehlers an.

Über den Zusammenhang zwischen Θ und L

Lemma

(i) Sei s^* minimierende Funktion von $\mathcal{E}(s) = \mathbb{P}[s(X) \neq Y]$ über alle messbaren Funktionen $s : \mathcal{X} \to \{-1, 1\}$. Dann gilt

$$\mathrm{E}[\ell(s^*)] = \min_{f:\mathcal{X}\to\mathbb{R} \text{ m.b.}} \mathrm{E}[\ell(f)].$$

Außerdem gilt, falls $f^* = \min_f \mathrm{E}[\ell(f)]$, dann ist $f^* = s^*$ fast sicher auf $\{\mathrm{P}[Y = 1 \mid X = x] \notin \{0, \frac{1}{2}, 1\}\}$.

(ii) Für jede P-messbare funktion f ist

$$\Theta(f, s^*) \leq L(f, f^*).$$

Beweis: (Blanchard et al., 2008, S.22).

Das Resultat von Blanchard et al. (2008)

Seien $\delta > 0$, $c \in \mathbb{R}$ und gelten weiterhin die folgenden Bedingungen:

- $\mathcal{H}_k$ sei separabel und $\forall x \in \mathcal{X} : k(x,x) \leq M^2 < \infty$
- Low Noise: $\forall x \in \mathcal{X} : |\eta(x) - \frac{1}{2}| \geq \eta_0$
- die Funktion $\gamma(n)$ sei ein Indikator für die „Kapazität“ des RKHS
- φ sei nicht fallende Funktion auf $\mathbb{R}^+$ mit $\varphi(0) = 0$ und $\varphi(x) \geq x$ für $x \geq \frac{1}{2}$

Dann gilt für $\Lambda_n > 0$ mit $\Lambda_n \geq c\left(\gamma(n) + \frac{\log(\delta^{-1}\log n)\vee 1}{n}\right)$ und

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}_k} \left(\frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \Lambda_n \varphi(M\|f\|_k) \right),$$

dass

$$\mathbb{P}\left[L(\hat{f}, s^*) \leq 2 \inf_{f \in \mathcal{H}_k} \left[L(f, s^*) + 2\Lambda_n \varphi(2M\|f\|_k) \right] + 4\Lambda_n (2\varphi(2) + \frac{c}{\eta_0}) \right] \geq 1 - \delta.$$

SVM als *Model Selection via penalization*

SVM_0 als *Penalization Method*

$$\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \Lambda_n \|f\|_k^2.$$

$$=$$

SVM_0 als *Model Selection via penalization*

$$\min_{R \in \mathbb{R}} \left\{ \min_{f : \|f\|_k \leq R} \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i f(X_i))_+ + \Lambda_n R^2 \right\}$$

Interpretation:

- Modelle: $f \in \mathcal{B}(R) := \{f \in \mathcal{H}_k \mid \|f\|_k \leq R\}$ im Norm-Ball mit Radius R
- Maß der Modell-Komplexität (bestraft durch penalization): Radius R

Definition

Eine Funktion $\psi : [0, \infty) \to [0, \infty)$ heiße *sub-root*, falls sie nicht negativ, nicht fallend und $r \mapsto \psi(r)/\sqrt{r}$ nicht steigend in $r > 0$ ist.

Sub-root Funktionen haben die folgende Eigenschaft:

Lemma

Sei $\psi : [0, \infty) \to [0, \infty)$ eine sub-root Funktion. Dann ist die Funktion stetig auf $[0, \infty)$ und die Fixpunktgleichung $\psi(r) = r$ hat eine eindeutige positive Lösung. Schreiben wir r^* für diese Lösung, so gilt für alle $r > 0$, dass $r \geq \psi(r) \Leftrightarrow r^* \leq r$.

Theorem

Sei $\mathcal{F} \subset L^2(\mathrm{P})$ eine Klasse messbarer Funktionen und weiterhin

1. $\exists b \in \mathbb{R}\ \forall f \in \mathcal{F} : \mathrm{P}\, f - f \leq b$
2. $w(f) : \mathcal{F} \to \mathbb{R}_{\geq 0}$ *mit* $\mathrm{Var}[f] \leq w(f)$
3. ϕ *sub-root,* $D > 0$, r^* *eindeutige Lösung von* $\phi(r) = r/D$
4. *es halte die folgende Aussage:*

$$\mathbb{E}\left[0 \vee \left(\sup_{f \in \mathcal{F} : w(f) \leq r} (\mathrm{P} - \mathrm{P}_n) f\right)\right] \leq \phi(r), \text{ für alle } r \geq r^*.$$

Dann gilt für alle $x > 0$ und alle $K > D/7$:

$$\mathbb{P}\left[\forall f \in \mathcal{F} : \mathrm{P}\, f - \mathrm{P}_n\, f \leq \frac{1}{K} w(f) + \frac{50K}{D^2} r^* + \frac{(K + 9b)x}{n}\right] \geq 1 - e^{-x}.$$

Beweis: (Blanchard et al., 2008, S.24 - 27).

Skizze des Beweises

Wähle:

- $\mathcal{M}$ eine abzählbare Indexmenge und $(R_m)_{m\in\mathcal{M}}$ eine passende diskretisierung von $\mathbb{R}$
- $\mathcal{F} = \mathcal{F}_{m,g_0} = \{\ell(g) - \ell(g_0) \mid g \in \mathcal{B}(R_m)\}$
- Pseudometrik d auf $L^2(\mathrm{P}_X)$ mit $d(g,g') := \mathrm{E}[(\ell(g) - \ell(g'))^2]$
- $w(f) = \min\{d(g,g_0) \mid g \in \mathcal{B}(R_m), f = \ell(g) - \ell(g_0)\}$

Dann halten (1), (2) von oben und es gibt (ϕ_m, D, r^*) aus (3), so dass (4) hält.
Sei weiterhin $g^* \in \arg\min_{g\in L^2(\mathrm{P}_X)} \mathrm{E}[\ell(g)]$, dann gibt es $C_m > 0$, so dass

$$\forall g \in \mathcal{B}(R_m) : d^2(g,g^*) \leq C_m L(g,g^*).$$

Beweis: (Blanchard et al., 2008, S.30 - 36).
Gelte nun für $\tilde{m} \in \mathcal{M}$, eine penalty Funktion $\mathrm{pen}(m)$ und positive Zahlen $(\rho_m)_{m\in\mathcal{M}}$, dass

$$\mathrm{P}_n\,\ell(\tilde{g}) + \mathrm{pen}(\tilde{m}) \leq \inf_{m\in\mathcal{M}} \left\{ \inf_{g\in\mathcal{B}(R_m)} (\mathrm{P}_n\,\ell(g) + \mathrm{pen}(m) + \rho_m) \right\}.$$

Skizze des Beweises - Zerlegung des relative average loss

Für alle $m \in \mathcal{M}$ und $g_m \in \mathcal{B}(R_m)$ gilt dann:

$$\begin{aligned}
& L(\tilde{g}, g^*) - L(g_m, g^*) \\
&= \mathrm{E}[\ell(\tilde{g})] - \mathrm{E}[\ell(g_m)] \\
&= \mathrm{E}_n[\ell(\tilde{g})] - \mathrm{E}_n[\ell(g_m)] + (\mathrm{P} - \mathrm{P}_n)(\ell(\tilde{g}) - \ell(g_m)) \\
&= \mathrm{E}_n[\ell(\tilde{g})] - \mathrm{E}_n[\ell(g_m)] + (\mathrm{P} - \mathrm{P}_n)(\ell(\tilde{g}) - \ell(u_{\tilde{m}})) + (\mathrm{P} - \mathrm{P}_n)(\ell(u_{\tilde{m}}) - \ell(g_m))
\end{aligned}$$

Nach Wahl von $\tilde{g}$:

$$\mathrm{E}_n[\ell(\tilde{g})] - \mathrm{E}_n[\ell(g_m)] \leq \mathrm{pen}(m) + \rho_m - \mathrm{pen}(\tilde{m})$$

Nach Theorem zu *localized uniform control* gilt in Wahrscheinlichkeit (abh. von ξ):

$$(\mathrm{P} - \mathrm{P}_n)(\ell(\tilde{g}) - \ell(u_{\tilde{m}})) \leq A\, d^2(\tilde{g}, g^*) + B\, r^*_{\tilde{m}} + O'_{\tilde{m}}(n, \xi)$$

Mit Hilfe einer Bernstein Ungleichung gilt in Wahrscheinlichkeit (abh. von ξ):

$$(\mathrm{P} - \mathrm{P}_n)(\ell(u_{\tilde{m}}) - \ell(g_m)) \leq D\, d^2(\tilde{g}, g^*) + E\, d^2(g_m, g^*) + O''_{\tilde{m}}(n, \xi) + O_m(n, \xi)$$

Skizze des Beweises - Zerlegung des relative average loss

Zusammenfassend gilt in Wahrscheinlichkeit (abh. von ξ) für alle $m \in \mathcal{M}$ und $g_m \in \mathcal{B}(R_m)$:

$$\begin{aligned} L(\tilde{g}, g^*) - L(g_m, g^*) \leq \operatorname{pen}(m) + \rho_m - \operatorname{pen}(\tilde{m}) \\ + (A+D)d^2(\tilde{g}, g^*) + E\, d^2(g_m, g^*) + O_m(n, \xi) + B\, r^*_{\tilde{m}} + O_{\tilde{m}}(n, \xi) \end{aligned}$$

Sei $\operatorname{pen}(m)$ so gewählt, dass $\operatorname{pen}(m) \geq B\, r^*_m + O_m(n, \xi)$, dann ist

$$\begin{aligned} \ldots &\leq (A+D)d^2(\tilde{g}, g^*) + E\, d^2(g_m, g^*) + 2\operatorname{pen}(m) + \rho_m \\ &\leq C_{\tilde{m}}(A+D)L(\tilde{g}, g^*) + C_m E\, L(g_m, g^*) + 2\operatorname{pen}(m) + \rho_m \end{aligned}$$

Umsortierung führt zu folgender Ungleichung in Wahrscheinlichkeit:

$$L(\tilde{g}, g^*) \leq K \inf_{m \in \mathcal{M}} \left\{ \inf_{g \in \mathcal{B}(R_m)} L(g, g^*) + 2\operatorname{pen}(m) + \rho_m \right\}.$$

Und Abschätzung der Ungleichung für Bälle mit Radii in $\mathbb{R}$ mit geschickter Diskretisierung $(R_m)_{m \in \mathcal{M}}$ führt zur Behauptung. □

Quellen und Verweise

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36(2):489–531, 04 2008. doi: 10.1214/009053607000000839. URL `https://doi.org/10.1214/009053607000000839`.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996. ISBN 978-1-4612-0711-5.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009. URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

D. Sejdinovic and A. Gretton. What is an rkhs? 2013. URL `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/RKHS_Notes1.pdf`.

Vielen Dank für Ihre Aufmerksamkeit!

Ein Theorem zur *model selection via penalization*

Sei $\ell : \mathfrak{G} \to L^2(P)$ [wobei $\mathfrak{G} \subset L^2(P)$] eine loss Funktion ist und nehme an, es existiere $g^* \in \arg\min_{g \in \mathfrak{G}} \mathrm{E}[\ell(g)]$. Sei $(\mathcal{G}_m)_{m \in \mathcal{M}}, \mathcal{G}_m \subset \mathfrak{G}$ eine abzählbare Kollektion von Funktionenklassen und nehme an, es existiere das folgende:

- eine Pseudometrik d auf $\mathfrak{G}$;
- eine Folge von sub-root Funktionen $(\phi_m), m \in \mathcal{M}$;
- zwei positive Folgen (b_m) und $(C_m), m \in \mathcal{M}$;

so dass,

$$\begin{aligned} \forall m \in \mathcal{M}, \forall g \in \mathcal{G}_m : \|\ell(g)\|_\infty &\leq b_m; \\ \forall g, g' \in \mathfrak{G} : \mathrm{Var}(\ell(g) - \ell(g')) &\leq d^2(g, g'); \\ \forall m \in \mathcal{M}, \forall g \in \mathcal{G}_m : d^2(g, g^*) &\leq C_m L(g, g^*); \end{aligned}$$

und mit r_m^* als Lösung zu $\phi_m(r) = r/C_m$,

$$\forall m \in \mathcal{M}, \forall g_0 \in \mathcal{G}_m, \forall r \geq r_m^* :$$

$$\mathrm{E}\left[\sup_{g \in \mathcal{G}_m, d^2(g, g_0) \leq r} (P - P_n)(\ell(g) - \ell(g_0))\right] \leq \phi_m(r).$$

Ein Theorem zur *model selection via penalization*

Theorem

Seien $\xi > 0$ und $K > 1$ reelle Konstanten und sei $\mathrm{pen}(m)$ eine penalty Funktion nach unten beschränkt durch eine Schranke $\mathcal{S}(m, \xi, K, n)$ (siehe Blanchard et al. (2008) für die Definition von $\mathcal{S}$). Seien außerdem $(\rho_m)_{M \in \mathcal{M}}$ positive reelle Zahlen und $\tilde{g}$ ein ρ_m-approximate penalized minimum empirical loss estimator über die Familie $\mathcal{G}_m$ zu der oben definierten penalty Funktion $\mathrm{pen}(m)$, so dass

$$\exists \tilde{m} \in \mathcal{M} : \tilde{g} \in \mathcal{G}_{\tilde{m}} \text{ und} \tag{1}$$

$$P_n \ell(\tilde{g}) + \mathrm{pen}(\tilde{m}) \leq \inf_{m \in \mathcal{M}} \inf_{g \in \mathcal{G}_m} (P_n \ell(g) + \mathrm{pen}(m) + \rho_m); \tag{2}$$

dann gilt:

$$\mathrm{P}\left[L(\tilde{g}, g^*) \leq \frac{K + 1/5}{K - 1} \inf_{m \in .\mathcal{M}} \left(\inf_{g \in \mathcal{G}_m} L(g, g^*) + 2\,\mathrm{pen}(m) + \rho_m \right) \right] > 1 - e^{-\xi}. \tag{3}$$