33) Durch die folgende Tabelle ist ein Trainingsdatensatz $\mathscr{T}$ zur Schätzung des monatlichen Stromverbrauchs t (in kWh) in Abhängigkeit von der Wohnfläche x_1 (in m^2) und der Anzahl der Bewohner x_2 gegeben:

Wohnung n	Wohnfläche x_{n1}	Bewohner x_{n2}	t_n
1	40	1	120
2	50	1	135
3	60	2	170
4	70	2	190
5	90	3	240
6	100	4	280

Der Stromverbrauch soll durch einen Regressionsbaum vorhergesagt werden, wobei als Verlustfunktion der quadratische Verlust

$$\sum_{(\mathbf{x}_n,t_n)\in\mathscr{T}} (t_n - \widehat{y}(\mathbf{x}_n))^2$$

verwendet wird. Runden Sie bei den Rechnungen auf eine Nachkommastelle.

a) Führen Sie eine erste Zerlegung entlang der Variablen Wohnfläch x_1 durch. Betrachten Sie hierbei die Schwellenwerte $t_1^* \in \{45, 55, 65, 80, 95\}$ und ermitteln Sie jeweils den resultierenden quadratischen Verlust. Welche Zerlegung minimiert den quadratischen Verlust?

b) Verzweigen Sie nun den rechten Teilbereich (d.h. die Beobachtungen mit $x_1 > t_1^*$) für den optimalen Schwellenwert aus Teilaufgabe a) weiter entlang der Variablen Anzahl der Bewohner x_2. Bestimmen Sie für die Schwellenwerte $t_2^* \in \{2{,}5; 3{,}5\}$ erneut den quadratischen Verlust und wählen Sie die beste Zerlegung.

c) Geben Sie den nach den Zerlegungen in den Aufgabenteilen a) und b) resultierenden Regressionsbaum (d.h. Wurzelknoten, innere Knoten, Endknoten und deren Prognosen $\widehat{y}_R$) an.

d) Berechnen Sie den Gesamtverlust des Baums nach beiden Zerlegungen und vergleichen Sie ihn mit dem Verlust des ungesplitteten Baums (nur ein Knoten). Wie stark verbessert sich die Anpassung?

e) Diskutieren Sie qualitativ, warum eine weitere Zerlegung des linken Teilbereichs hier nicht sinnvoll ist (Stichwort: Overfitting vs. lokale Varianzreduktion).

37) Betrachtet wird ein Regressionsproblem mit dem Trainingsdatensatz

$$\mathscr{T} = \Big\{ (x_n, t_n) \in \mathbb{R}^2 : n = 1, \ldots, N \Big\},$$

dem das wahre datengenerierende Modell

$$t = \beta_0 + \beta x + \varepsilon$$

mit $\mathbb{E}[\varepsilon] = 0$ und $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$ zugrunde liegt. Durch Ziehen mit Zurücklegen werden B Bootstrap-Trainingsdatensätze

$$\mathscr{T}_1^*, \mathscr{T}_2^*, \ldots, \mathscr{T}_B^*$$

vom Umfang N erzeugt und damit anschließend B instabile Basislerner $\widehat{y}_b(x)$ für $b = 1, \ldots, B$ ermittelt (z.B. Regressionsbäume), welche die lineare Regressionsfunktion $f(x) := \beta_0 + \beta x$ jeweils mit einem systematischen Bias und Varianz

$$b(x) := \mathbb{E}\big[\widehat{y}_b(x)\big] - f(x) \qquad \text{und} \qquad \mathrm{Var}\big(\widehat{y}_b(x)\big) = \sigma_y^2$$

schätzen. Zwischen zwei beliebigen verschiedenen Basislerner $\widehat{y}_b(x)$ und $\widehat{y}_b(x)$ bestehe die Korrelation $\rho \in [0, 1)$.

Übungsaufgaben

a) Zeigen Sie für die Bagging-Prognose

$$\widehat{y}_{\text{bag}}(x) := \frac{1}{B}\sum_{b=1}^{B}\widehat{y}_b(x)$$

gilt

$$\mathbb{E}\left[\widehat{y}_{\text{bag}}(x)\right] = \beta_0 + \beta x + b(x) \quad \text{und} \quad \text{Var}\left(\widehat{y}_{\text{bag}}(x)\right) = \sigma_y^2\left(\rho + \frac{1-\rho}{B}\right).$$

Interpretieren Sie das Ergebnis.

b) Weisen Sie nach für das erwartete Risiko (erwarteter quadratischer Fehler) $R\left(\widehat{y}_{\text{bag}}(x)\right) = \mathbb{E}\left[(\widehat{y}_{\text{bag}}(x) - t)^2\right]$ gilt:

$$R\left(\widehat{y}_{\text{bag}}(x)\right) = \underbrace{\sigma_y^2\left(\rho + \frac{1-\rho}{B}\right)}_{\text{Var}(\widehat{y}_{\text{bag}}(\mathbf{x}))} + \underbrace{b(x)^2}_{\text{Bias}^2} + \underbrace{\sigma_\varepsilon^2}_{\text{Datenrauschen}}$$

c) Zeige Sie, dass Bagging das erwartete Risiko gegenüber einem einzelnen Basislerner genau dann verringert, wenn

$$\sigma_y^2(1-\rho)\left(1 - \frac{1}{B}\right) > 0$$

gilt. Interpretieren Sie ferner das Ergebnis.

Übungsaufgaben

38) Betrachtet wird ein Random Forest mit $p = 10$ Inputvariablen von denen $r = 2$ relevant sind. Bei der rekursiven binären Zerlegung des Inputraums $\mathscr{D} \subseteq \mathbb{R}^p$ wird bei jeder Verzweigung zufällig ohne Zurücklegen eine Menge aus $m \leq p$ Inputvariablen ausgewählt, die für die Verzweigung in Frage kommen (sog. feature bagging).

   a) Bestimme die Wahrscheinlichkeit, mindestens eine der r relevanten Variablen in der ausgewählten m-elementigen Teilmenge zu haben.
   b) Berechne die Wahrscheinlichkeit für $r \in \{2, 3, 5, 10\}$.
   c) Interpretiere Sie den Einfluss von m auf Bias und Varianz sowie auf die dekorrelierende Wirkung der randomisierten Variablenauswahl.

41) Gegeben sei ein Trainingsdatensatz

$$\mathscr{T} = \{(\mathbf{x}_n, t_n) \in \mathbb{R}^p \times \{-1,1\} : n = 1,\ldots,5\}$$

für ein binäres Klassifikationsproblems mit $t_n \in \{-1,1\}$ und die folgende Tabelle:

Beobachtung n	t_n	Lerner $k_b^A(\mathbf{x}_n)$	Lerner $k_b^B(\mathbf{x}_n)$
1	$+1$	$+1$	-1
2	$+1$	-1	$+1$
3	-1	-1	-1
4	-1	$+1$	-1
5	$+1$	$+1$	$+1$

Es sei $y_{b-1}(\mathbf{x}) \equiv 0$ (Start).

a) Führe eine AdaBoost-Iteration b durch und gib den verstärkten Klassifikator $y_b(\mathbf{x}) = y_{b-1}(\mathbf{x}) + \alpha_b k_b(\mathbf{x})$ an.

b) Berechnen Sie die neuen Gewichte $w_n^{(b+1)} = \exp\big(-t_n y_b(\mathbf{x}_n)\big)$ für $n = 1,\ldots,5$.