

1.3.4) Parametrische & nicht-parametrische Modelle

gesucht ist eine Funktion \hat{f} , so dass $Y \approx \hat{f}(X)$ für jede Beobachtung (X, Y)

Zwei Arten von Modellen:

① Parametrische Modelle

Zwei Schritte: 1) Mache Annahmen über die funktionale Struktur von f

Bsp.: $f(X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ \leftarrow lineares Modell

Modellparameter: $\beta_0, \beta_1, \dots, \beta_p$

2) Benütze die Trainingsdaten um β_0, \dots, β_p so festzulegen, dass die Residuen $Y - \hat{Y}$ minimal werden (z.B. LQ-Schätzung)

☞ Das allgemeine Problem wird auf das Finden optimaler Modellparameter reduziert \Rightarrow wenig Flexibilität

Das ist deutlich einfacher, als eine komplett beliebige Funktion f anzupassen!



Nachteil dabei:

wir müssen f a priori festlegen.

Dieses f passt normalerweise nur bedingt zum darunterliegenden, unbekannten wahren Modell M_{true} .

2) Nicht-parametrische Modelle

Nicht-parametrische Verfahren machen keine expliziten Annahmen über die funktionale Form von $f \Rightarrow$ sehr flexibel

Es wird nur versucht möglichst nahe an den Daten zu liegen
 \rightsquigarrow komplexere Strukturen können abgebildet werden



Nachteil dabei: das Problem wird nicht auf eine vorgegebene Modellklasse (mit kleiner Zahl zu schätzender Parameter) reduziert
 \Rightarrow man benötigt viel mehr Daten um ein gutes Modell schätzen zu können.

Bsp.: Splines

einzige Annahme: die Modellfunktion soll "glatt" sein.

Wie "glatt" muss vorher festgelegt werden

Bemerkung: nicht-parametrische Modelle neigen i. A. dazu sich zu sehr an die Trainingsdaten anzupassen (overfitting) und schlecht zu generalisieren.
Dann muss man entsprechend gegensteuern ...

Allgemein gilt:

- flexible Modelle sind anpassungsfähig, aber wenig interpretierbar \Rightarrow schlecht für Inferenz
- unflexible Modelle machen oft schlechtere Prognosen, können aber besser interpretiert werden



Gibt es allgemeine Prinzipien um die Modellgüte (und damit auch die Modelleigenschaften) bewerten zu können?

1.4) Bewertung der Modellgüte



welches Modell (in welcher Konfiguration) ist am besten für meine Aufgabe geeignet ??

... nicht leicht zu entscheiden!

Man muß oft mehrere Modellierungsversuche starten und die Ergebnisse vergleichen.

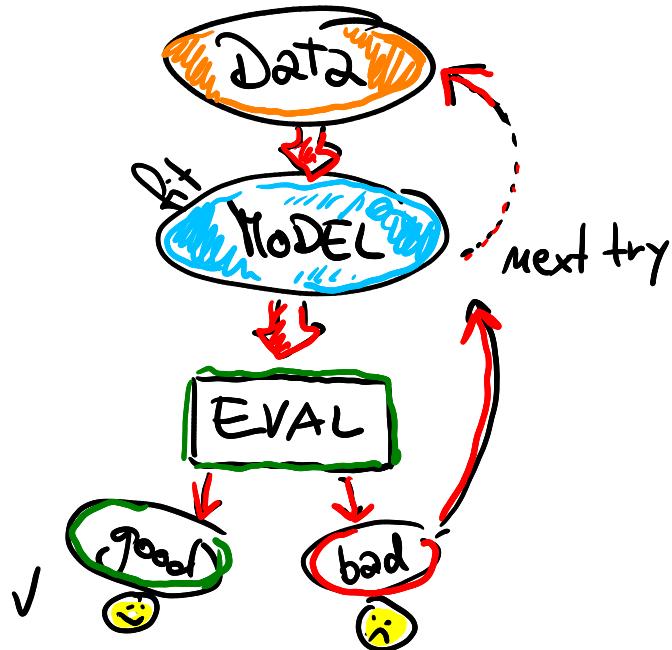


man braucht geeignete

Gütemasse



Diese Metriken sehen je nach Art der Aufgabe sehr unterschiedlich aus!



1.4.1) Überlegungen für Regressionsprobleme

Eines der gebräuchlichsten Fehlermaße bei Regressionsproblemen ist der mittlere quadratische Fehler ($MSE = \text{Mean Squared Error}$)

Def.: MSE

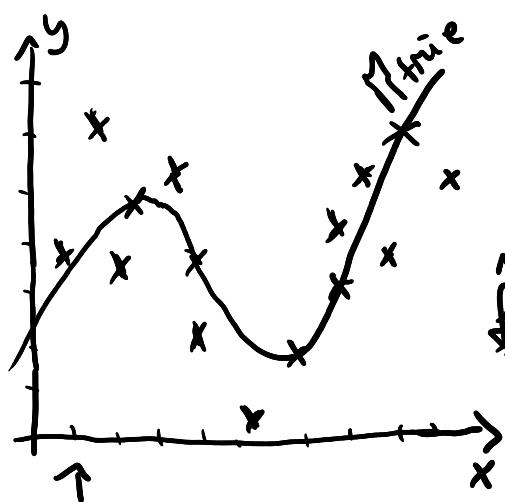
Seien y_i und x_i ($i=1, \dots, n$) Beobachtungen von Y und X , sowie \hat{f} eine geschätzte Transferfunktion. Dann gilt

$$MSE := \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$$

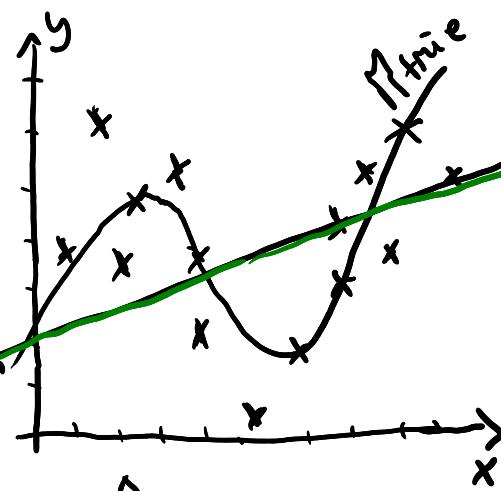
Wird man den Fehler nur auf den Trainingsdaten, so spricht man vom Trainingsfehler.

Wirklich interessant ist aber eigentlich der Fehler auf der Testmenge (Testfehler). Die würde ja nicht zum Lernen benötigt, die Zielgröße ist aber bekannt
 ⇒ simuliert den "Erfolg"

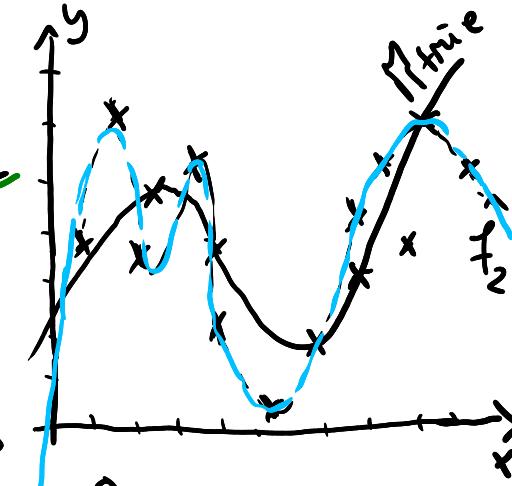
Unabhängig von benützten Modelltyp lässt sich die folgende Beobachtung machen:



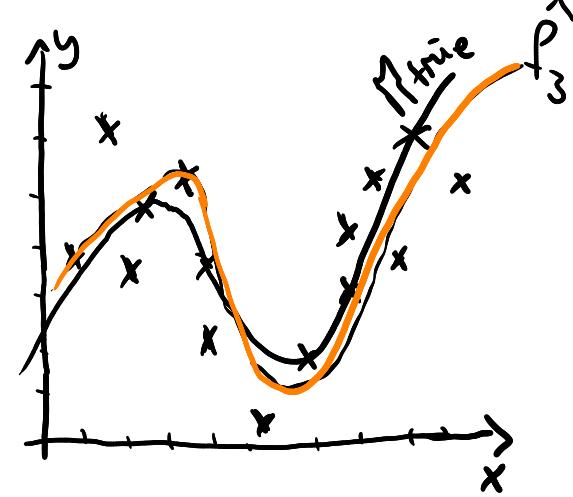
- f_{true} ist das wahre Modell. \times Datenpunkte



- \hat{f}_2 ist eine lineare Transferfunktion



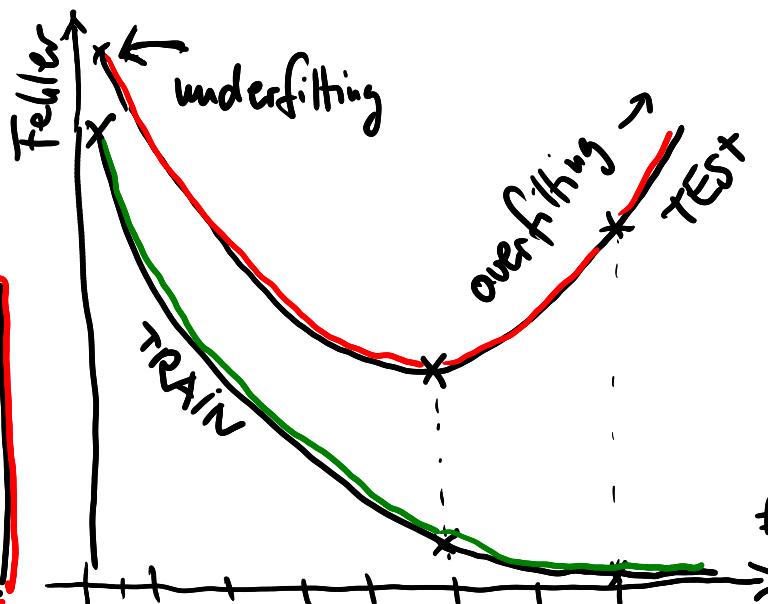
- \hat{f}_3 hat richtig niedrige Komplexität \Rightarrow Paßt top zu den Daten



- \hat{f}_3 ist nicht so komplex erfasst aber den datenerzeugenden Prozeß

Für den Fehler bedeutet das:

wir brauchen die Testmenge zur Ermittlung der notwendigen Komplexität



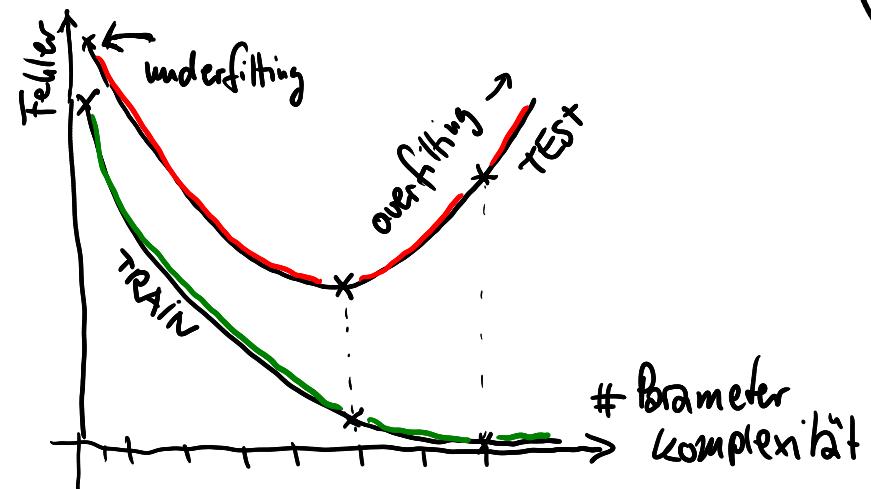
→ hoher Train\Teufehler
⇒ undershifting

→ kleiner Trainings- aber großer Teufehler ⇒ overfitting

Parameter
Komplexität

1.4.2) Das "Bias-Variance"-Dilemma

Wie lässt sich erklären, dass beim Testfehler von ganz unterschiedlichen Verfahren mit wachsender Komplexität immer eine ähnliche Struktur zu erkennen ist ??

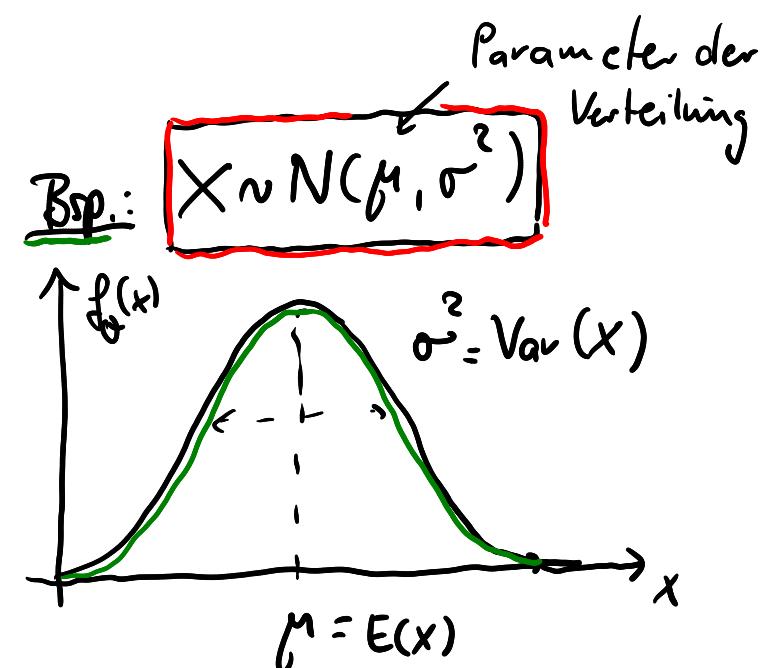


Das muss einen Grund haben 😊

Vorüberlegungen aus der Statistik

Wir betrachten eine ZV X mit einer Verteilung $f_Q(x)$, $Q \hat{=} \text{Parameter der Verteilung}$.

ZIEL → Schätze die unbekannten Parameter aus den Daten (= Realisierungen der Zufallsvariable X)



normalverteilte Größe mit Lageparameter μ , Streuungspar. σ^2

schen gesehen:

Für die Schätzung von θ aus einer Stichprobe braucht man eine Schätzfunktion

Def.: Sei θ ein unbekannter Parameter einer Verteilung $F_\theta(x)$ und $x = (x_1, \dots, x_n)$ eine Stichprobe. Dann heißt

$T_\theta(x) = T_\theta(x_1, \dots, x_n)$ eine Schätzfunktion von θ

Für die Realisierung der Stichprobe $x = (x_1, \dots, x_n)$ ergibt sich der konkrete Schätzwert $\hat{\theta}$ als

$$\hat{\theta} = T_\theta(x_1, \dots, x_n) = t(x_1, \dots, x_n)$$

Idealweise ist die Schätzfunktion $T_\theta(x)$ erwartungstreu d.h.

$$E(T_\theta(x)) = \theta$$

Das ist eine TOP-Eigenschaft 😊, denn

⇒ keine systematische Über- oder Unterschätzung!

⇒ die Schätzfunktion ist sinnvoll konstruiert!

⚠ ist aber leider nicht immer der Fall
⇒ Verzerrung!

die Verzerrung nicht erwartungstreuer Schätzer messen wir mit dem

 Def.: Bias

$$\text{bias}(T_\theta(x)) = E(T_\theta(x)) - \theta$$

← Abweichung des Erwartungswertes der Schätzung von θ

ABER: Der Fehler beim Schätzen hängt nicht nur vom bias ab sondern leider auch von der Varianz der Schätzfunktion $V(T_\theta(x))$. Das wird im MSE zusammengefasst:

$$\text{MSE}(T_\theta(x)) = E((T_\theta(x) - \theta)^2) = V(T_\theta(x)) + \text{bias}(T_\theta(x))^2$$

Für erwartungstreue $T_\theta(x)$ gilt also

$$\text{MSE}(T_\theta(x)) = V(T_\theta(x))$$

wg. $\text{bias} = 0$!

 Def.:

$$\sigma_{T_\theta} = \sqrt{V(T_\theta(x))}$$

heißt Standardfehler der Schätzung



Was heißt das für unsere Modelle?

betrachte Trainingsdaten (x_i, y_i) mit

$$y_i = f(x_i) + \varepsilon$$

$$E(\varepsilon) = 0$$

$$\text{Var}(\varepsilon) = \sigma^2$$

\Rightarrow geschätzte Funktion $\hat{f}(x_i) = \hat{y}_i$ mit

$$\text{bias}(\hat{f}(x)) = E((\hat{f}(x) - f(x))^2)$$

Dann beobachtet man Residuen $y_i - \hat{f}(x_i)$ und $\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$

Finde ein \hat{f} , so dass $(y - \hat{f}(x))^2$ minimal ist!

Es gilt: $E((y - \hat{f}(x))^2) = E[y^2 + \hat{f}^2(x) - 2y\hat{f}(x)] \stackrel{\text{R.R.}}{=} E(y^2) + E(\hat{f}^2(x)) - 2E(y\hat{f}(x))$

wegen $\star E(y^2) = \text{Var}(y) + E(y)^2$ gilt:

$$E((y - \hat{f}(x))^2) = \text{Var}(y) + E(y)^2 + \text{Var}(\hat{f}(x)) + E(\hat{f}(x))^2 - 2E(y) \cdot E(\hat{f}(x))$$

aber unser Modell hat die Eigenschaft $f = E(y) \Rightarrow$

$$E[(y - \hat{f}(x))^2] = \text{Var}(y) + \underbrace{\hat{f}^2}_{\text{Formel}} + \text{Var}(\hat{f}(x)) + \underbrace{E(\hat{f}(x))^2 - 2 \cdot \text{bias}(\hat{f}(x))}_{\text{bias}}$$

bias
Formel

(45)

$$= \text{Var}(y) + \text{Var}(\hat{f}(x)) + (\hat{f}(x) - E(\hat{f}(x)))^2 \quad \text{Bias!!}$$

$$y = f(x) + \varepsilon$$

betrachte $\text{Var}(y)$: $\text{Var}(y) = E((y - E(y))^2) = E((y - f(x))^2) =$

$$= E((f(x) + \varepsilon - f(x))^2) = E(\varepsilon^2) = \text{Var}(\varepsilon) + E(\varepsilon)^2$$

$$\Rightarrow \boxed{\text{Var}(y) = \text{Var}(\varepsilon) = \sigma^2} \quad \leftarrow \text{hängt nur von der Streuung der Störgröße ab !!}$$

$$\Rightarrow \boxed{E[(y - \hat{f}(x))^2] = \text{Var}(\hat{f}(x)) + [\text{bias}(\hat{f}(x))]^2 + \text{Var}(\varepsilon)} \quad \text{nicht reduzierbarer Fehler !!}$$

Fazit → Der erwartete Testfehler besteht aus dem nicht-reduzierbaren Fehler $\text{Var}(\varepsilon) = \sigma^2$ und der reduzierbaren Fehler zerfällt in 2 Teile:

$$\text{Var}(\hat{f}(x)) \quad \text{und} \quad (\text{bias}(\hat{f}(x)))^2$$

Um einen optimal niedrigen Fehler zu bekommen brauchen wir also ein Lernverfahren, das gleichzeitig niedrige Varianz und niedrigen bias aufweist !!

Was heißt das?

→ niedrige Varianz heißt, das Ergebnis ändert sich nur wenig, wenn andere Trainingsdaten verwendet werden!

- Flexible Verfahren haben normalerweise eine hohe Varianz (hängen an Daten)

→ niedriger bias heißt, das Modell ist komplex genug um die Struktur abzubilden und keine systematischen Fehler zu machen!

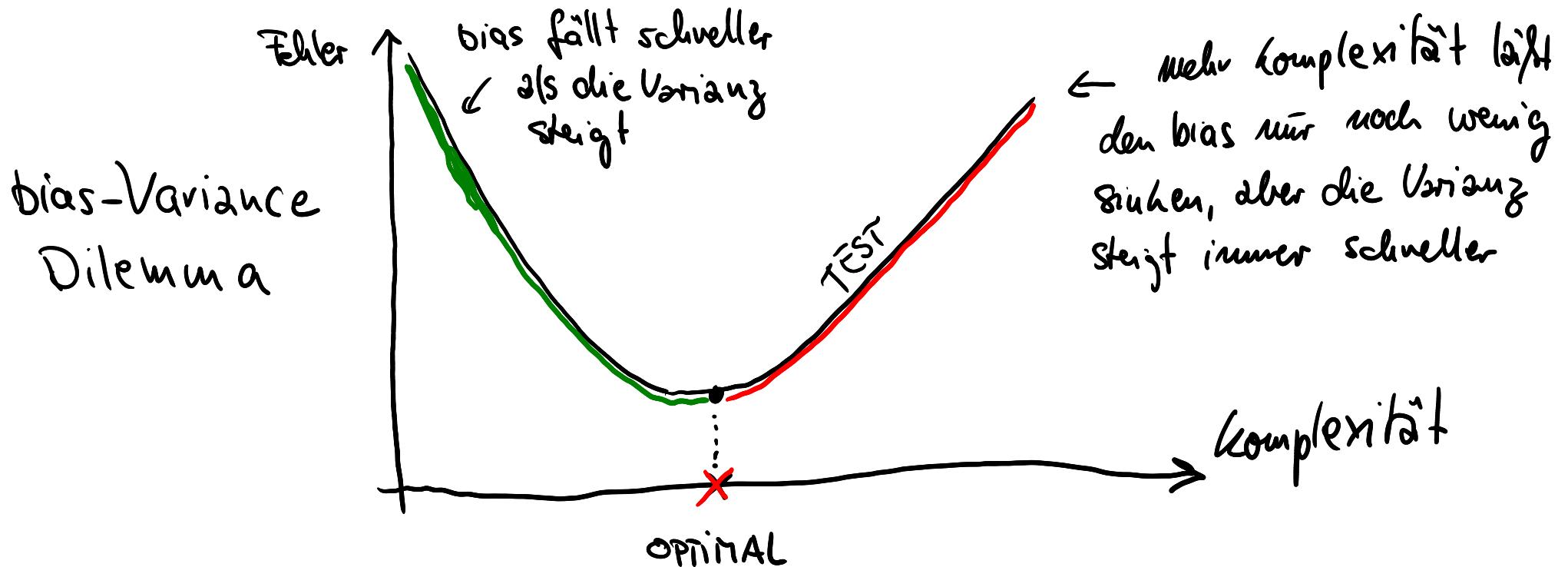
- unflexible Methoden (z.B. linear) haben einen großen bias

⇒ je flexibler (komplexer) das Modell ist, desto niedriger ist der bias
ABER: die Varianz steigt mit wachsender Komplexität





Das Zusammenspiel von Bias und Varianz bestimmt den Verlauf des Testfehlers!



- Bsp.:
- f linear
 - ⇒ LinReg hat bias = 0
 - ⇒ flexible Methoden haben großen Fehler

- f hochgradig nicht-linear
- ⇒ LinReg hat riesigen bias
- ⇒ flexible Verfahren schneiden besser ab

1.4.3) Überlegungen für Klassifikationsprobleme

(52)

Aufgabe: Ordne mit Hilfe der Features X_1, \dots, X_p korrekte Labels Y zu

Die y_i ($i=1 \dots, n$) sind nicht mehr numerisch sondern qualitativ
Als Gütemaß für ein Modell \hat{f} kann man dann die Trefferquote
("Hit Rate") oder die Fehlerquote ("Error Rate") verwenden ...

Def: Error Rate

Betrachte Trainingsdaten (x_i, y_i) ($i=1 \dots, n$) mit $\hat{f}(x_i) = \hat{y}_i$ ($i=1 \dots, n$)

Dann heißt

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$$

Fehlerrate von \hat{f}

$I(\cdot) = 1$ für
 $y_i \neq \hat{y}_i$ 0 sonst

Die Fehlerrate gibt den Anteil der Fehlklassifizierungen in der
Trainingsmenge (bzw. Testmenge) wieder.



Wie gut kann man für gegebene Daten maximal werden? Also: gibt es auch hier einen irreduziblen Fehler??

→ **BAYES - CLASSIFIER**

Def.: Bayes - Classifier

Ordne einer Beobachtung (x_0, y_0) aus der Teststichprobe mit dem Feature-Vektor x_0 diejenige Klasse $j \geq 1$, für die gilt:

$$P(Y=j | X=x_0)$$

ist maximal

↗ BAYES - CLASSIFIER

bedingte Wahrscheinlichkeit $Y=j$ zu beobachten unter der Bedingung, dass $X=x_0$ gilt

Es lässt sich tatsächlich zeigen, dass der Bayes-Classifier den minimalen Testfehler besitzt, denn

Fehler bei $X=x_0$ ist

$$1 - \max_j (P(Y=j | X=x_0))$$

\Rightarrow Die allgemeine Bayes-Fehlerrate

$$1 - E(\max_j (P(Y=j | X)))$$

entspricht dem irreduziblen Fehler bei der Regression!

MW über alle möglichen Werte von X

ABER ! Der Bayes-Classifier lässt sich NICHT (exakt) berechnen!

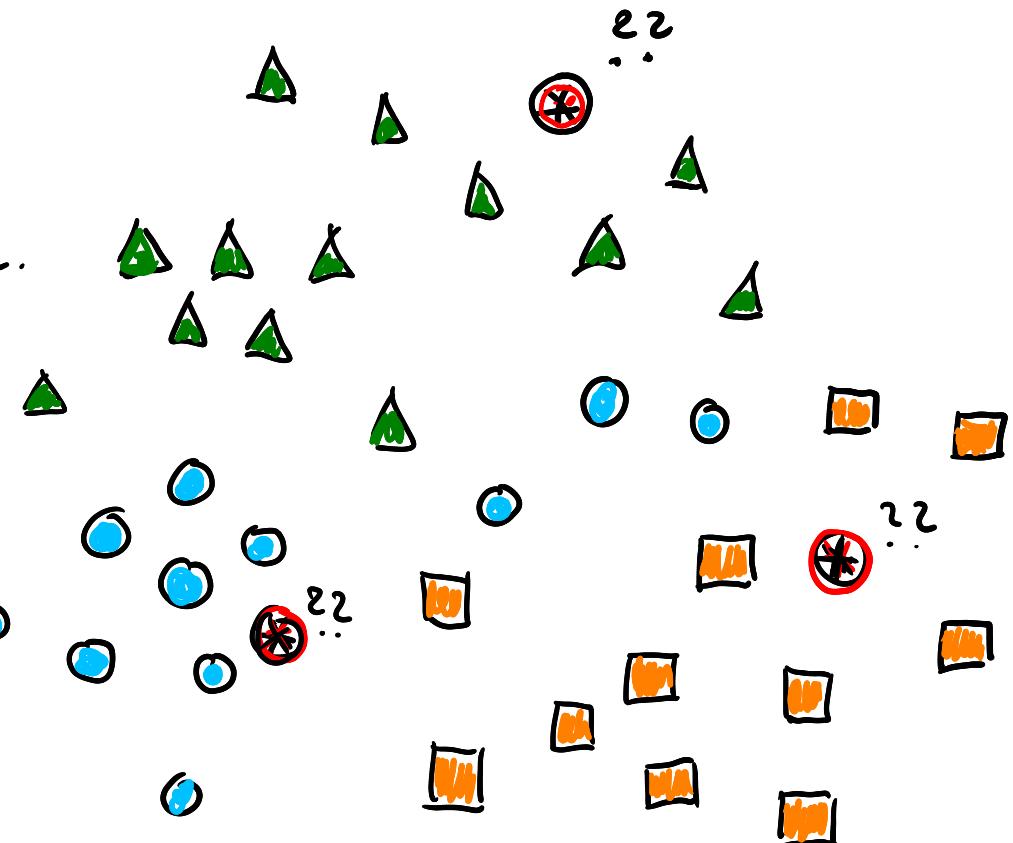
Wir kennen ja die wahre bedingte Verteilung für $Y|X$ nicht.

\Rightarrow viele Methoden zur Klassifikation versuchen die bedingte Verteilung $Y|X$ zu SCHÄTZEN

Bsp.: KNN

K-nearest neighbor - Verfahren

- entscheide Dich für ein $k=1, 2, 3, \dots$
- finde die k nächsten Nachbarn für X^*
- die Mehrheit der Nachbarn entscheidet über die Klasse von X^*



Bemerkung:

Das "Bias-Variance"-Dilemma gilt auch für die Komplexität von Klassifikationsverfahren.

Nur die Metriken für den Loss sind anders...

\triangle Dreieck
 \square Quadrat
 \circ Kreis