Behandelte Themen

0. „Motivation“: Lernen in Statistik und Biologie

1. Überblick über statistische Datenmodellierungs-Verfahren

2. Das lineare Modell (Regression)

3. Perceptron und Multilagen-Perceptron (Funktionsapproximation)

4. Selbstorganisierende Merkmalskarten (Dichteschätzung)

5. Lernen von Datenmodellen

- Approximation: Bayes‘sches Schließen, MAP, ML
- Maximum Likelihood Schätzung und Fehlerminimierung
- Generalisierung und Regularisierung
- Optimierungsverfahren

6. Bayesianische Netze (Dichteschätzung und Funktionsapproximation)

7. Kern-Trick und Support Vector Machine

Lernen von Datenmodellen

Approximation

- **Bayesianische Philosophie und Bayes‘sches Schließen**
- **Bayesianische Datenmodellierung**
- **Spezialfall: Maximum-Likelihood-Schätzung**

Ziel maschinellen Lernens (=statistische Inferenz)

Gegeben:

- **Man verfügt nur über einen Datensatz** $D = \left\{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)} \right\}$ **als Beispiel**
- **Statistische Datenmodelle verfügen über freie Parameter** $\mathbf{w}$

Ziel:

- **Gute modellhafte Beschreibung der zugrundeliegenden statistischen Struktur.... Also**
- **Gute Beschreibung des Datensatzes** D **(=> Optimierung der Parameter** $\mathbf{w}$**) ... Aber auch**
- **Gute Generalisierung, d.h. Beschreibung neuer Datensätze (=> Regularisierung)**

Bayesianische Philosophie und Bayes‘sches Schließen

Wiederholung:

Frequentistische Philosophie:

- **Der Datenpunkt ist eine Stichprobe aus einer existierenden „wahren“ Verteilung** $p(x)$
- **=>** $p(x)$ **kann als Grenzwert der relativen Häufigkeit interpretiert werden**
- **„Wahrscheinlichkeit als Prozentsatz“**
- **Typisches Beispiel: Wahrscheinlichkeit, 6 zu würfeln**

Bayesianische Philosophie:

- **Nur der Datenpunkt** x_0 **existiert, es gibt keine zugrundeliegende Verteilung**
- **„Wahrscheinlichkeit als Glaube (Belief) des Eintretens eines Sachverhalts“**
- **Basiert allein auf Vorwissen und auf den beobachteten Daten**
- **Typisches Beispiel: Wahrscheinlichkeit, den nächsten Marathon zu gewinnen**

Motivation:

- **Es gilt, die Daten zu erklären.**
- **Es gibt kein einzelnes zugrundeliegendes Modell**
- **Bestimme Wahrscheinlichkeitsverteilung über die Modelle $\mathbf{w}$, gegeben**
  -- den Datensatz D
  -- unser Vertrauen (Vorwissen, „Prior Belief“) in die Modelle

Prinzip: Bayes‘sches Gesetz

$$p(\mathbf{w} \mid D) = \frac{1}{p(D)} p(D \mid \mathbf{w}) p(\mathbf{w})$$

$p(\mathbf{w})$ **= a Priori-Wissen bzw. Vertrauen (Belief) in ein Modell ohne Daten („prior“)**

$p(D \mid \mathbf{w})$ **= Wahrscheinlichkeit der Daten im Lichte des Modells $\mathbf{w}$ („likelihood“)**

$p(\mathbf{w} \mid D)$ **= a Posteriori Wahrsch. (Belief) in ein Modell geg. Daten („posterior“)**

$p(D)$ **= Evidenz („evidence“), bewertet Güte der Modellfamilie (vgl. „Hyperparameter“)**

Beispiel zum Bayes'schen Schließen:

Ein Labortest gibt mit Unsicherheit behaftete Auskunft über einen Krebstyp

-- Man weiß, dass 0.8 % der Bevölkerung diesen Krebs entwickeln.

-- Der Test reagiert in 98% der Fälle positiv (+), wenn der Patient Krebs hat.

-- Der Test reagiert in 97% der Fälle negativ (-), wenn der Proband gesund ist.

Frage: Wie sicher hat der Proband Krebs, wenn das Testergebnis bekannt ist?

Vor dem Test:

$$p(krebs) = 0.008$$
$$p(gesund) = 0.992$$

$$p(+ \mid krebs) = 0.98$$
$$p(- \mid krebs) = 0.02$$
$$p(+ \mid gesund) = 0.03$$
$$p(- \mid gesund) = 0.97$$

$$p(+) = p(+ \mid krebs)\, p(krebs) + p(+ \mid gesund)\, p(gesund)$$
$$= 0.98 \cdot 0.008 + 0.03 \cdot 0.992$$
$$= 0.0376$$
$$p(-) = 0.9624$$

Nach dem Test:

$$p(krebs \mid +) = p(+ \mid krebs)\, p(krebs) / p(+) = 0.98 \cdot 0.008 / 0.0376 = 0.209$$
$$p(krebs \mid -) = p(- \mid krebs)\, p(krebs) / p(-) = 0.02 \cdot 0.008 / 0.9624 = 0.00016$$

-- Eintreffen eines Datenpunktes verändert die Schätzung

-- Positiver Test erhöht die Krebschance nur moderat

Bayesianische Datenmodellierung

Bayes'sche Dichteschätzung:

- **Benutze alle Modelle zur Schätzung der Wahrscheinlichkeit für neuen Datenpunkt x:**

$$p(\mathbf{x} \mid D) = \int p(\mathbf{x} \mid \mathbf{w}) p(\mathbf{w} \mid D) d\mathbf{w}$$

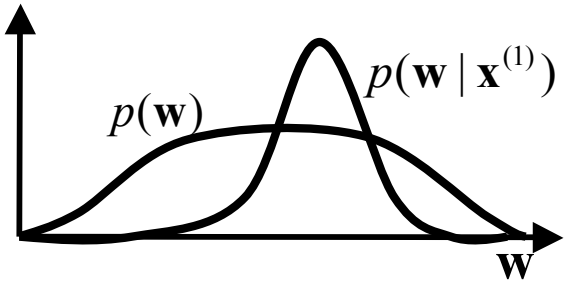
Bayes'sche Regression:

- **Marginalisiere über Modelle und bilde Erwartungswert über Outputs:**

$$\hat{y}(x \mid D) = \int y \int p_n(y - f(x \mid \mathbf{w})) p(\mathbf{w} \mid D) d\mathbf{w} dy$$

Iteratives Lernen von Bayes-Schätzern:

$$D = \left\{ \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \right\}$$



- **Vor dem ersten Datenpunkt:** $p(\mathbf{w})$
- **Nach dem ersten Datenpunkt:** $p(\mathbf{w} \mid \mathbf{x}^{(1)}) \propto p(\mathbf{x}^{(1)} \mid \mathbf{w}) p(\mathbf{w})$
- **Nach dem zweiten Datenpunkt:** $p(\mathbf{w} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \propto p(\mathbf{x}^{(2)} \mid \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}^{(1)})$

. . . $p(\mathbf{w} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \propto p(\mathbf{x}^{(2)}, \mathbf{x}^{(1)} \mid \mathbf{w}) p(\mathbf{w}) = p(\mathbf{x}^{(2)} \mid \mathbf{w}) p(\mathbf{x}^{(1)} \mid \mathbf{w}) p(\mathbf{w}) \propto p(\mathbf{x}^{(2)} \mid \mathbf{w}) p(\mathbf{w} \mid \mathbf{x}^{(1)})$

Likelihood-Funktion und konjugierter Prior:

- **Beobachtung: Posterior im m-ten Schritt ist Prior im $m+1$. Schritt**
- **Wünschenswert: Funktionelle Form von Prior und Posterior sollen übereinstimmen**
- **Geg: Likelihood-Funktion $p(D \mid \mathbf{w})$**
- **Def: Prior $p(\mathbf{w})$ ist zur Likelihood-Funktion konjugiert, wenn $p(\mathbf{w})$, $p(\mathbf{w} \mid D)$ dieselbe funktionelle Form haben**
- **Bsp:**

Likelihood (Fkt. von x, param. durch w)		**konjugierter Prior** (Fkt. von w)	
Gauss	$p(x \mid \mu) \propto \exp(-(x-\mu)^2 / 2\sigma^2)$	**Gauss**	$p(\mu) \propto \exp(-(\mu-\alpha_1)^2 / 2\alpha_2{}^2)$
Exp.	$p(x \mid w) = w\exp(-wx)$	**Gamma**	$p(w) \propto w^{\alpha_1} \exp(-\alpha_2 w)$
Binomial	$p(k \mid w) \propto w^k (1-w)^{M-k}$	**Beta**	$p(w) \propto w^{\alpha_1 - 1}(1-w)^{\alpha_2 - 1}$

- α_1, α_2 **Bsp. für Hyperparameter des Prior,** σ **Bsp. für Likelihood-Hyperparameter**

Hierarchische Bayes-Modelle:

- **Hierarchisches Modell: Schätzung der Hyperparameter aus der Evidenz**

$$p(\mathbf{w}) \equiv p(\mathbf{w} \mid \alpha), \quad p(D \mid \mathbf{w}) \equiv p(D \mid \mathbf{w}, \beta)$$

α= Prior-Hyperparameter, β= Likelihood-Hyperparameter. Die Evidenz wird

$$p(D) \equiv p(D \mid \alpha, \beta) = \int p(D \mid \mathbf{w}, \beta)\, p(\mathbf{w} \mid \alpha) d\mathbf{w}$$

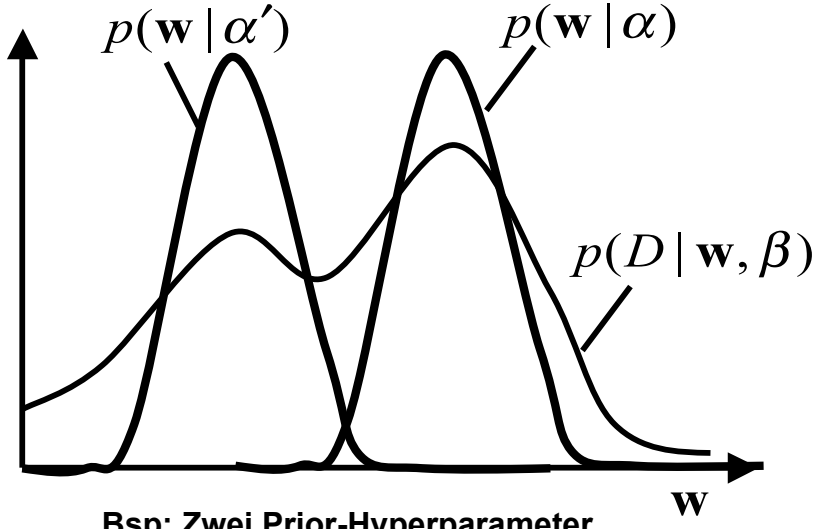
- **„Evidenz“ = Evidenz von α, β im Lichte der Daten**
  = Fähigkeit der gesamten Modellklasse, beschrieben durch α, β,
  die Daten zu erklären

- **Bayes‘sche Schätzung der Hyperparameter:**

$$p(\alpha, \beta \mid D) \equiv \frac{p(D \mid \alpha, \beta)\, p(\alpha, \beta)}{p(D)}$$

$p(\alpha, \beta)$ **= Hyperprior**

- **Hierarchische Modellierung:**
  Evidenz Level k = Likelihood Level k+1



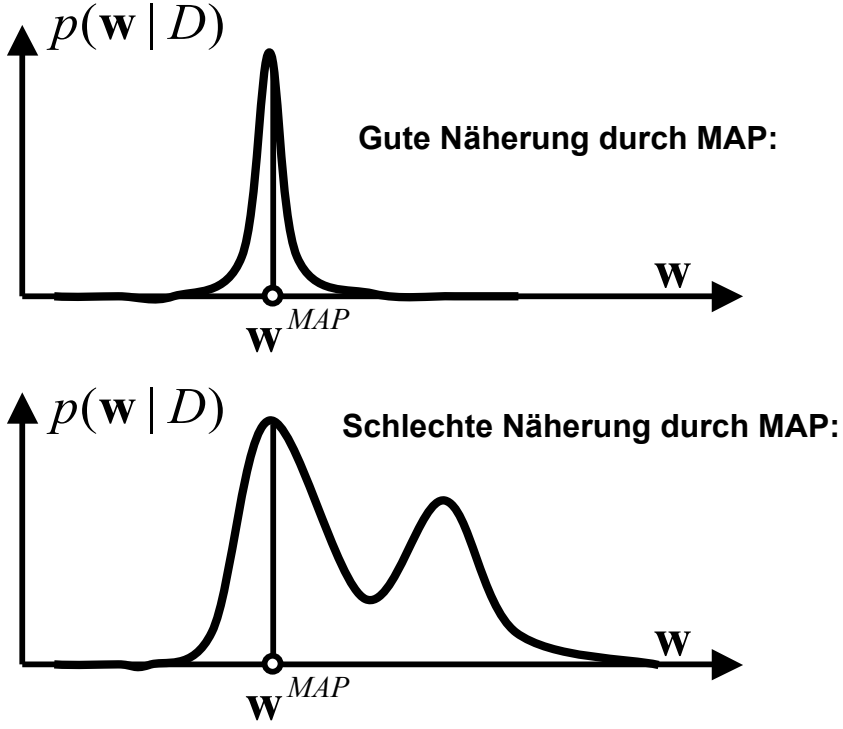
Bsp: Zwei Prior-Hyperparameter

Maximum a Posteriori (MAP) Schätzer:

- **Marginalisierung über Modelle oft schwierig**
- **Näherung: Verwende das Modell mit der größten a-posteriori-Wahrscheinlichkeit**

$$\mathbf{w}^{MAP} = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid D)$$

- **Gute Näherung bei konzentriertem, symmetrischem posterior (=> viele Daten)**
- **Übergang zu frequentistischer Sicht (relative Häufigkeiten, ein wahres Modell)**



MAP Lernen: Maximiere posterior $p(D \mid \mathbf{w}) p(\mathbf{w}) = \max$

Äquivalent : Minimiere -log posterior $-\ln p(D \mid \mathbf{w}) - \ln p(\mathbf{w}) + const = \min$

⇧ ⇧

Später: Risikominimierung Regularisierung

- **Nur ein Modell, aber das muss man finden! => „Optimierung“**

Maximum-Likelihood-Parameterschätzung

Ziel:

- **Maximiere Wahrscheinlichkeit, dass die Daten aus dem Modell erzeugt wurden => Maximum Likelihood**
- **Bem: Maximum-Likelihood (ML) entspricht MAP ohne Vorwissen**

Prinzip:

- **Likelihood:** $p(D \mid \mathbf{w}) = p(\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\} \mid \mathbf{w})$
- **Max. Likelihood:** $\mathbf{w}^{ML} = \arg\max_{\mathbf{w}} p(D \mid \mathbf{w})$
- **Äquivalent:** $\mathbf{w}^{ML} = \arg\min_{\mathbf{w}} -\ln p(\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\} \mid \mathbf{w}) =: \arg\min_{\mathbf{w}} L(\mathbf{w})$

  $L(\mathbf{w})$ **wird auch als negative log-likelihood oder „Likelihood-Funktion“ bezeichnet**

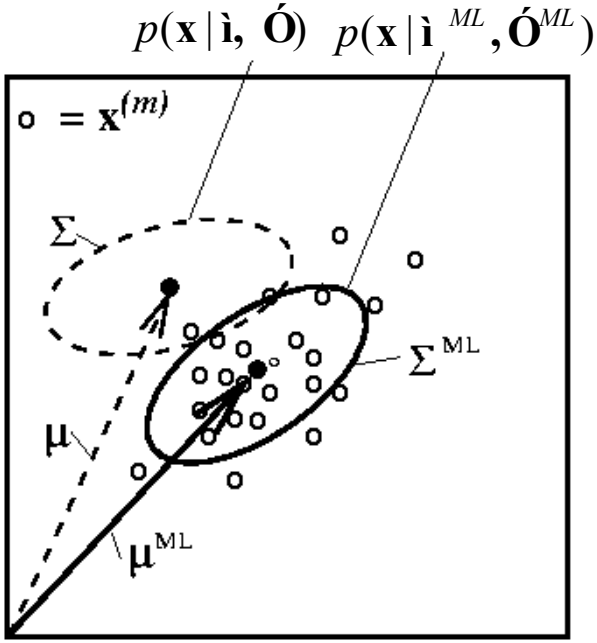
- **Für iid („independent, identically distributed“) gezogene Datenpunkte sind die Likelihood-Funktionen verschiedener Datenpunkte unabhängig:**

$$L(\mathbf{w}) = -\ln p\left(\left\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}\right\} \mid \mathbf{w}\right) = -\ln \prod_{m=1}^{M} p(\mathbf{x}^{(m)} \mid \mathbf{w}) = -\sum_{m=1}^{M} \ln p(\mathbf{x}^{(m)} \mid \mathbf{w})$$

Bsp: ML-Schätzung einer Gauss-Verteilung (parametrische Dichteschätzung):

$$\varphi(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}(\det\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{m=1}^{M}\frac{1}{2}(\mathbf{x}^{(m)}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(m)}-\boldsymbol{\mu}) + \sum_{m=1}^{M}\frac{1}{2}\ln\det\boldsymbol{\Sigma} + c$$



Eindimensional: $$L(\mu,\sigma^2) = \sum_{m=1}^{M}\frac{(x^{(m)}-\mu)^2}{2\sigma^2} + M\ln\sigma + c$$

ML-Schätzung für μ: Verschwindende Ableitung:

$$\frac{\partial}{\partial\mu}L(\mu,\sigma^2)\stackrel{!}{=}0 = -\frac{1}{\sigma^2}\sum_{m=1}^{M}(x^{(m)}-\mu) \Rightarrow \boxed{\mu^{ML} = \frac{1}{M}\sum_{m=1}^{M}x^{(m)}}$$

ML-Schätzung für σ:

$$\frac{\partial}{\partial\sigma}L(\mu,\sigma^2)\stackrel{!}{=}0 = -\sum_{m=1}^{M}\frac{(x^{(m)}-\mu^{ML})^2}{\sigma^3} + \frac{M}{\sigma} \Rightarrow \boxed{\sigma^{2ML} = \frac{1}{M}\sum_{m=1}^{M}(x^{(m)}-\mu^{ML})^2}$$

Bem: Bei Gauss-verteilten Daten mißt L gerade die mittlere quadratische Abweichung der Daten vom Mittel („quadratischer Fehler“, siehe später)