Statistik

Vorlesung 9 - Parameterschätzung Teil 2:
Maximum Likelihood Estimate (MLE) und Maximum A Posteriori (MAP)

Prof. Dr. Sandra Eisenreich

Hochschule Landshut

Einführung: MLE und MAP

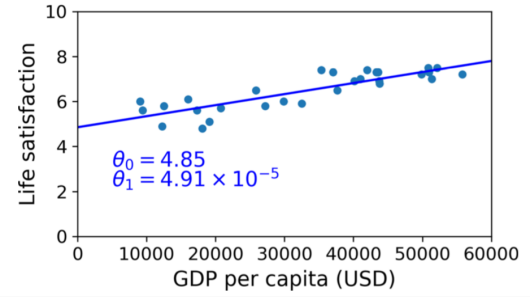
Motivation

Bisher haben wir bei unseren Schätzern "geraten", und dann nachgewiesen, dass es gute Schätzer sind. Wie jedoch kann man auch in komplizierteren Fällen einen Schätzer konstruieren?

Beispiel: Einfache Lineare Regression

Die folgenden Daten D sind Stichproben einer Gaußschen Normalverteilung mit Erwartungswert = die eingezeichnete Gerade mit Dichte

$$N(y|f(x), \sigma^2).$$



Geradengleichung: $f(x) = \theta_0 + \theta_1 \cdot x$,
Parameter:

- θ_0 = y-Abschnitt bei $x = 0$
- θ_1 = Steigung der Geraden

Aufgabe: Bestimme Parameter θ_0, θ_1, die zu den Daten passen! Was heißt das?

Schätzmethoden für Parameter

Maximum Likelihood Schätzmethode

Stehen verschiedene Modelle P_θ zur Konkurrenz, so halte bei vorliegenden Daten D dasjenige Modell für das glaubwürdigste, unter welchem die beobachteten Daten die größte Wahrscheinlichkeit des Auftretens besitzen, das heißt:

$$\text{maximiere } P(D|\theta)$$

Maximum A Posteriori Schätzmethode

Stehen verschiedene Modelle P_θ zur Konkurrenz, so halte bei vorliegenden Daten D dasjenige Modell für das glaubwürdigste,welches unter den beobachteten Daten die größte Wahrscheinlichkeit des Auftretens besitzt, das heißt:

$$\text{maximiere } P(\theta|D)$$

Beispiel: Linda, die Bankangestellte

Daten D: Linda ist 31 Jahre alt, single, gibt ihre Meinung offen und direkt kund, und ist sehr klug. Sie hat Philosophie studiert. Als Studentin hat sie sich sehr zu den Themen von Diskriminierung und sozialer Ungerechtigkeit engagiert und an Anti-Atomkraft-Demonstrationen teilgenommen.

Welche der folgenden "Parameter" sind Ihrer Meinung nach besser?

Parameter 1: Linda ist Bankangestellte.
Parameter 2: Linda ist Bankangestellte und engagiert sich für die Gleichberechtigung von Frauen.

Beispiel: Maximum Likelihood

A="Eine Person ist Bankangestellte"
B="Eine Person engagiert sich für Gleichberechtigung"
D="Eine Person ist 31 Jahre alt, single, etc. "
Parameter 1 $= A$, Parameter 2 $= A \cap B$
$A \cap B \subset A \Rightarrow P(\text{Parameter 2}) = P(A \cap B) \leq P(A) = P(\text{Parameter 1})$

Maximum Likelihood Estimate: maximiere $P(D|\text{Parameter})$, also: welche Parameter machen die Daten (31, sagt ihre Meinung, engagiert sich gegen Diskriminierung, etc.) wahrscheinlichker?

$$p(D|\text{Parameter 1}) = P(D|A) = \frac{P(D \cap A)}{P(A)} < \frac{p(D \cap A)}{p(A \cap B)} = p(D|\text{Parameter 2})$$

Der Maximum Likelihood Schätzwert für die Parameter wäre also: Parameter 2!

Beispiel: Maximum A Posteriori

Maximum A Posteriori Schätzwert: maximiere die Wahrscheinlichkeit der Parameter, wenn die Daten D gegeben sind, also: maximiere
also: ist Parameter 1 oder Parameter 2 wahrscheinlicher, wenn wir die Daten über Linda wissen (31, sagt ihre Meinung, engagiert sich gegen Diskriminierung, etc.)?

Wir wissen bereits:

$$P(\text{Parameter 2}) < P(\text{Parameter 1})$$

$$P(\text{Parameter 2}|D) = \frac{P(\text{Parameter 2} \cap D)}{P(D)} < \frac{P(\text{Parameter 1} \cap D)}{P(D)} = P(\text{Parameter 1}|D)$$

Der Maximum A Posteriori Schätzwert wären also Parameter 1!

Diskrete und stetige Verteilungen

Im Folgenden wollen wir sowohl im diskreten, wie auch im stetigen Setting "Wahrscheinlichkeiten" maximieren. Damit meinen wir

- im diskreten Fall die Verteilungen $P_\theta(X = x)$ abhängig von unbekannten Parameter(n) θ.
- im stetigen Fall die Wahrscheinlichkeitsdichte $f_\theta(x)$, welche der Verteilungsfunktion $P_\theta(X \leq x)$ zugrunde liegt.

Regel für das folgende Kapitel: Damit wir nicht jedes Mal zwischen beiden Fällen unterscheiden müssen, schreiben wir immer P_θ, meinen aber die Dichte f_θ, wenn wir im stetigen Fall sind.

Maximum Likelihood Estimate (MLE)

Motivation: Likelihood-Funktion

gegeben: Stichprobe $X = x$

gesucht: $\widehat{\theta}$, so dass $P_{\widehat{\theta}}(X = x)$ maximal

Idee: Betrachte $L_x(\theta) := P_\theta(X = x)$ als Funktion in θ (genannt Likelihood-Funktion) und bestimme das Maximum dieser Funktion: $\text{argmax}_\theta L_x(\theta)$

Ist $f(\theta)$ eine Funktion, so schreibt man für das $\widehat{\theta}$, bei dem f sein Maximum (bzw. Minimum) annimmt:

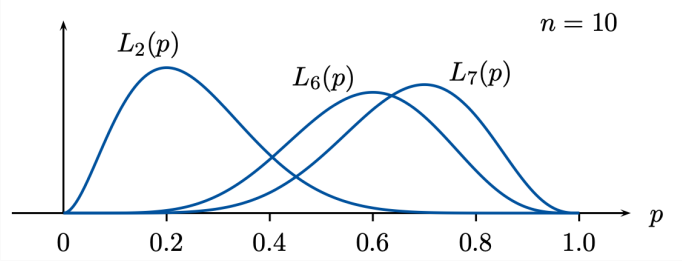
$$\widehat{\theta} \quad = \quad \text{argmax}_\theta f(\theta) \quad (\text{bzw. } \text{argmin}_\theta f(\theta))$$

Beispiel: Werfen einer gezinkten Münze

10x Werfen einer gezinkten Münze, $\theta = p = P(\text{Kopf})$. $X =$ "#Kopf".

$$L_x(\theta) = P_\theta(X = x) = \binom{10}{x} \theta^x (1-\theta)^{10-x}$$

Für unterschiedliche Stichproben bekommen wir andere Likelihood-Funktionen:



Gesucht: Die Maximalstelle, hier z.B. $\theta = \frac{2}{10}$ für $x = 2$, $\theta = \frac{6}{10}$ für $x = 6$...

Maximum-Likelihood-Schätzer

Sei $(\mathcal{X}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell, $x = (x_1, \ldots, x_n) \in \mathcal{X}$ eine Stichprobe.

(a) Die Likelihood-Funktion für θ zur Beobachtung $X = x$ ist:

$$L_x(\theta) \quad := \quad P_\theta(X = x)$$

(b) Existiert ein $\hat{\theta}(x) \in \Theta$ mit $\hat{\theta}(x) = \text{argmax}_\theta L_x(\theta)$, also

$$L_x(\hat{\theta}(x)) = \sup_{\theta\in\Theta} L_x(\theta)$$

so heißt $\hat{\theta}(x)$ Maximum-Likelihood-Schätzwert/-Estimate (MLE) für θ zu x.

(c) Der Schätzer, der jeder Stichprobe $x \in \mathcal{X}$ den dazugehörigen Maximum Likelihood Schätzwert $\hat{\theta}(x)$ zuordnet, heißt Maximum-Likelihood-Schätzer.

Motivation: Loglikelihood-Funktion

Für eine iid Stichprobe $x = (x_1, \ldots, x_n)$ gilt:

$$L_x(\theta) \quad = \quad P_\theta(X_1 = x_1, \ldots, X_n = x_n) = \prod_{j=1}^{n} P_\theta(X_j = x_j).$$

gesucht: Maximalstelle.

Wie? ableiten und Nullstelle suchen!

Problem: Für große n ist die Ableitung von dem Produkt schwer zu berechnen. (Kettenregel...)

Trick zur leichteren Berechnung des Maximums

Ansatz: Eine Summe ist leichter abzuleiten als ein Produkt. Wie kann man aus dem Produkt eine Summe machen?

Idee: Der Logarithmus tut das: $\log(a \cdot b) = \log a + \log b$.

Der Logarithmus ist streng monoton steigend, das heißt:

- je größer y ist, umso größer ist $\log(y)$.
- je größer $y = L_x(\theta)$ ist, umso größer ist $\log L_x(\theta)$
- Das heißt: wenn $y = L_x(\theta)$ sein Maximum erreicht, dann auch $\log L_x(\theta)$!

$\Rightarrow$ **Neues Ziel:** Maximiere

$$LL_x(\theta) := \log L_x(\theta) = \log \left(\prod_{j=1}^{n} P_\theta(X_j = x_j) \right) = \sum_{j=1}^{n} \log P_\theta(X_j = x_j)$$

Loglikelihood-Funktion

- Die Loglikelihood-Funktion zu x ist $LL_x(\theta) := \log L_x(\theta) = \log P_\theta(X = x)$
- Für iid Stichproben $x = (x_1, \ldots, x_n)$ gilt:

$$LL_x(\theta) = \sum_{j=1}^{n} \log P_\theta(X_j = x_j).$$

- Negative LogLikelihood: $NLL_x(\theta) := -LL_x(\theta)$.

Für den Maximum-Likelihood-Schätzwert $\hat{\theta}$ zur Stichprobe x gilt

$$\hat{\theta} = \text{argmax}_\theta LL_x(\theta) = \text{argmin}_\theta NLL_x(\theta).$$

Vorgehen zum Ermitteln vom Maximum Likelihood Schätzwert

0. Welches Zufallsexperiment?
1. Welche Art der Wahrscheinlichkeitsverteilung (z.B. Binomialverteilung, Normalverteilung) mit noch unbekannten Parameter(n)?
2. Berechne die Loglikelihood Funktion $LL_x(\theta)$.

3a. Nur ein Parameter θ: $LL'(\theta) = 0$ liefert Extremstelle $\hat{p}$; falls $L''(\hat{p}) < 0$ ist $\hat{p}$ ein Maximum.

3b. Gibt es mehrere Parameter $\theta = (\theta_1, \ldots, \theta_m)$, so gehe ähnlich vor:

- Bestimme Parameter θ_i, indem man LL_x partiell nach θ_i ableitet (und die anderen Parameter wie Konstanten behandelt)
- Alle solchen Ableitungen $= 0$ setzen liefert m Gleichungen für $\theta_1, \ldots, \theta_m$
- Ob diese Extremstelle ein Maximum ist, erhält man mit mehr mathem. Theorie.

Beispiel: MLE für 10 Münzwürfe

Wir erinnern uns an das Beispiel mit den 10 Würfen mit einer gezinkten Münze. Angenommen, die Stichprobe ergibt $x = 2$ mal Kopf. Was ist das MLE für p?

Ergebnis: MLE für 10 Münzwürfe

0. Zufallsexperiment: Bernoulli
1. Wahrscheinlichkeitsverteilung: binomial mit $\theta = p = P(\text{Kopf}) \in (0,1)$ unbekannt
2. $L_2(p) = b_{10,p}(X = 2) = \binom{10}{2} p^2 \cdot (1-p)^8 = 45p^2 \cdot (1-p)^8 \Rightarrow LL_2(p) = \log\binom{10}{2} + 2\log p + 8 \cdot (1-p)$.

3a. $LL_2'(p) = \frac{2}{p} - \frac{8}{1-p}$, also gilt:

$$LL_2'(\hat{p}) = 0 \quad \Leftrightarrow \quad \frac{2}{p} = \frac{8}{1-p} \Leftrightarrow 1 - \hat{p} = 4\hat{p} \Leftrightarrow 1 = 5\hat{p} \Leftrightarrow \hat{p} = \frac{1}{5}$$

Ist $\hat{p} = 1/5$ ein Minimum? Berechne dazu

$$LL_2''(p) = -\frac{2}{p^2} - \frac{8}{(1-p)^2} < 0 \Rightarrow \text{Maximum}$$

$\hat{p} = 1/5$ ist also der Maximum-Likelihood-Schätzwert

Allgemein: MLE für Bernoulli-Experiment

Wir betrachten ein allgemeines n-stufiges Bernoulli-Experiment $(X_1, \ldots, X_n)$ mit Stichprobe $(x_1, \ldots, x_n)$ mit genau $x = x_1 + \ldots x_n$ Treffern.

Für ein n-stufiges Bernoulli-Experiment mit genau x Treffern ist der MLE die relative Trefferhäufigkeit

$$\hat{p} = \frac{x}{n}$$

und der MLE-Punktschätzer ist gegeben durch $T = \frac{S_n}{n}$.

Begründung: MLE für Bernoulli-Experiment

0. Zufallsexperiment: Bernoulli
1. Wahrscheinlichkeitsverteilung: binomial mit $\theta = p = P(\text{Kopf}) \in (0,1)$ unbekannt
2. $L_x(p) = b_{n,p}(X = x) = \binom{n}{x} p^x \cdot (1-p)^{n-x}$
   $LL_x(p) = \log \binom{n}{x} + x \cdot \log p + (n-x) \cdot \log(1-p)$
3. $LL'_x(p) = \frac{x}{p} - \frac{n-x}{1-p}$, also gilt:

$$LL'_x(p) = 0 \quad \Leftrightarrow \quad \frac{x}{p} = \frac{n-x}{1-p} \Leftrightarrow x(1-p) = (n-x)p \Leftrightarrow x = n \cdot p \Leftrightarrow \hat{p} = \frac{x}{n}$$

Man kann wie vorher nachrechnen dass hier $L''_x(\frac{x}{n}) < 0$, es ist also ein Maximum.

ML-Schätzung bei geometrischer Verteilung

Wir führen n viele iid mehrstufige Bernoulli-Experimente durch mit Parameter θ, und $X_i =$ "Versuch, bei dem erstmalig ein Treffer auftritt". Gegeben: Stichprobe $(x_1, \ldots, x_n)$. gesucht: die Trefferwahrscheinlichkeit θ eines Bernoulli-Experiments.

Beispiel: Wir führen n-mal das Experiment durch: 10x Werfen einer gezinkten Münze, $\theta = p = P(\text{Kopf})$. $X =$ "Experiment bei dem das erste Mal Kopf auftritt".

$$L_x(\theta) = P_\theta(X = x) = (1-p)^{x-1}p$$

Ergebnis: ML-Schätzung bei geometrischer Verteilung

0. Zufallsexperiment: n mehrstufige Bernoulli-Experimente X_i
1. Wahrscheinlichkeitsverteilung eines X_i: geometrisch mit $\theta = p \in [0, 1]$ unbekannt: $P(X_i = x_i) = p(1-p)^{x_i - 1}$
2. Die Likelihood Funktion ist $L_{(x_1,\ldots,x_n)}(p) = \log(\prod_{i=1}^{n} P(X_i = x_i))$, also ist die Loglikelihood-Funktion:

$$\begin{aligned} LL_{(x_1,\ldots,x_n)}(p) &= \sum_{i=1}^{n} \log(p(1-p)^{x_i-1}) = \sum_{i=1}^{n} (\log p + (x_i - 1)\log(1-p)) \\ &= n \log p + (x_1 + x_2 + \ldots + x_n - n) \cdot \log(1-p). \end{aligned}$$

Ergebnis: ML-Schätzung bei geometrischer Verteilung

3. $LL'_{(x_1,\ldots,x_n)}(p) = \frac{n}{p} - \frac{x_1+x_2+\ldots+x_n-n}{1-p}$, also gilt:

$$\begin{aligned} LL'_{(x_1,\ldots,x_n)}(p) = 0 \quad &\Leftrightarrow \quad \frac{n}{p} = \frac{x_1 + x_2 + \ldots + x_n - n}{1-p} \\ &\Leftrightarrow \quad n(1-p) = (x_1 + x_2 + \ldots + x_n - n)p \\ &\Leftrightarrow \quad 1 - p = \frac{\sum_{i=1}^n x_i - n}{n} p = \frac{\sum_{i=1}^n x_i}{n} p - p \\ &\Leftrightarrow \quad \frac{n}{\sum_{i=1}^n x_i} = p \end{aligned}$$

Man kann wieder nachrechnen, dass hier $LL''_{(x_1,\ldots,x_n)}(\frac{n}{\sum_{i=1}^n x_i}) < 0$, es ist also ein Maximum.

Für n mehrstufige Bernoulli-Experimente mit dem ersten Treffer bei $(x_1, \ldots, x_n)$ Versuchen ist der MLE $\hat{p} = \frac{n}{\sum_{i=1}^n x_i}$.

Maximum A Posteriori Schätzwert

Satz von Bayes und Maximum a Posteriori

Maximum Likelihood: maximiere $P_\theta(X = x) = P(X = x|\theta)$

Maximum a Posteriori: maximiere $P(\theta|X = x)$. Diese Größe nennt man posterior distribution/ a posteriori Verteilung.

Frage: Wie kann man $P(\theta|X = x)$ aus der Likelihood $P(X = x|\theta)$ berechnen? Für das Ereignis $\{X = x\}$ schreiben wir von nun an nur x.

Satz von Bayes! Danach gilt:

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$$

A-priori und a-posteriori Verteilungen

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)},$$

- $P(\theta|x)$ heißt a-posteriori Verteilung oder posterior distribution
- $P(x|\theta) =$ Likelihood
- $P(\theta)$ heißt a-priori Verteilung oder prior distribution
- $P(x)$ heißt marginal likelihood oder evidence.

Bemerkungen zu a-priori Verteilung und marginal likelihood

- $P(\theta)$ ist eine Annahme, welche Art Verteilung die Parameter haben.
- $P(x)$ ist die totale Wahrscheinlichkeit, die Stichprobe x zu erhalten (egal unter welcher Verteilung). Sie hängt nicht von θ ab, ist also eine Konstante.

Bayes für stetige und diskrete Verteilungen

Bedingte Wahrscheinlichkeit und die Formel von Bayes gelten auch für die Dichten von stetigen Verteilungen. Es gilt also

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$$

wenn wir weiterhin schreiben:

$$P = \begin{cases} \text{Verteilung} & \text{im diskreten Fall} \\ \text{Dichte} & \text{im stetigen Fall} \end{cases}$$

Beispiel: Werfen einer gezinkten Münze

10x Werfen einer gezinkten Münze, $\theta = p = P(\text{Kopf})$. $X =$ Anzahl von "Kopf", angenommen $x = 2$. Dann ist die a-posteriori-Wahrscheinlichkeit gegeben durch:

$$P(p|2) = \frac{P(2|p) \cdot P(p)}{P(2)},$$

wobei wir als a-priori Verteilung eine Normalverteilung mit Varianz 0.1 annehmen, also

$$P(p) = N(p|0, 0.1) = \frac{1}{\sqrt{0.2\pi}} \exp\left(-\frac{p^2}{0.2}\right)$$

Wie berechnen wir MAP?

Ergebnis: Werfen einer gezinkten Münze

- Likelihood $= P(2|p) = \binom{10}{2} p^2(1-p)^8$
- Wir nehmen als a-priori Verteilung eine Normalverteilung mit Varianz 0.1 an, also

$$P(p) = N(p|0, 0.1) = \frac{1}{\sqrt{0.2\pi}} \exp\left(-\frac{p^2}{0.2}\right)$$

- marginal likelihood $P(2) =$ konstant $=$ totale Wahrscheinlichkeit von 2-mal Kopf für alle möglichen Werte von p.

Gesucht: Die Maximalstelle von $f(p) = P(p|2)$, aber da $P(2) =$ konstant ist diese dieselbe wie die Maximalstelle von $F(p) = P(2|p) \cdot P(p)$.

Allgemein: Berechnung des MAP Schätzwerts

gesucht: der Parameter $\hat{\theta}$, so dass

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$$

maximal wird. Da $P(x)$ allerdings konstant ist, ist das gleichbedeutend dazu, dass

$$P(x|\theta) \cdot P(\theta) = L_x(\theta) \cdot P(\theta)$$

maximal wird. Nun verwenden wir denselben Trick wie bei MLE um aus Produkt eine Summe zu machen: Da der Logarithmus streng monoton steigend ist, ist das gleichbedeutend dazu, dass

$$\log(L_x(\theta) \cdot P(\theta)) = LL_x(\theta) + \log P(\theta)$$

maximal wird.

Maximum-A-Posteriori-Schätzer

Es seien $(\mathcal{X}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Modell, $x = (x_1, \ldots, x_n) \in \mathcal{X}$ eine Stichprobe. Gegeben sei außerdem eine prior distribution $P(\theta)$.

(a) Betrachte die Funktion $F_x(\theta) := LL_x(\theta) + \log P(\theta)$

(b) Existiert ein $\hat{\theta}_{MAP}(x) \in \Theta$ mit $\hat{\theta}_{MAP}(x) = \operatorname{argmax}_\theta F_x(\theta)$, also

$$F_x(\hat{\theta}_{MAP}(x)) = \sup_{\theta\in\Theta} F_x(\theta) \tag{1}$$

so ist $\hat{\theta}_{MAP}(x)$ ist die Maximalstelle von $P(\theta|x)$ und heißt Maximum-A-Posteriori-/MAP-Schätzwert für θ zu x.

(c) Der zugehörige Schätzer heißt Maximum-A-Posteriori-/MAP-Schätzer.

Vorgehen zum Ermitteln vom Maximum Likelihood Schätzwert

gegeben: eine a-priori Verteilung $P(\theta)$

0. Welches Zufallsexperiment?
1. Berechne die Loglikelihood Funktion $LL_x(\theta)$ und daraus $F_x(\theta) = LL_x(\theta) + \log P(\theta)$.
2. Weiter wie bei Maximum Likelihood, nur dass wir statt $LL_x(\theta)$ $F_x(\theta)$ maximieren.

Beispiel: Werfen einer gezinkten Münze

10x Werfen einer gezinkten Münze, $\theta = p = P(\text{Kopf})$. $X =$ "#Kopf".

Stichprobe bei einem Bernoulli-Experiment liefert $x = 2$.

Als a-priori-Verteilung nehmen wir eine Normalverteilung an:

$$P(p) := N(p|0, 0.1) = \frac{1}{\sqrt{0.2\pi}} \exp\left(-\frac{p^2}{0.2}\right)$$

Berechne den MAP-Schätzwert.

Ergebnis: Werfen einer gezinkten Münze

0. Bernoulli, $L_2(p) = \binom{10}{2} p^2(1-p)^8$
1. $F_2(p) = LL_2(p) + \log P(p) = LL_2(p) + const - \frac{p^2}{0.2}$
2. $F_2'(p) = \frac{2}{p} - 8\frac{1}{1-p} - 10p$ Man kann berechnen dass eine Nullstelle dieser Funktion gegeben ist durch $p \simeq 0,18$ und $F_2''(0.18) < 0$ (die anderen beiden Ergebnisse sind nicht zwischen 0 und 1 und ergeben keinen Sinn). Hier haben wir also ein anderes Ergebnis als bei Maximum Likelihood!

Der MAP-Schätzwert ist kleiner als der MLE-Schätzwert, wenn man als a-priori Verteilung die Normalverteilung nimmt! Das hat einen Grund, siehe nächste Seite.

Ausblick: MAP a-priori Normalverteilung

Es seien $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ ein statistisches Modell mit $\theta = (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$, $x = (x_1, \ldots, x_n) \in \mathcal{X}$ eine Stichprobe. Die prior distribution sei $P(\theta) = \mathcal{N}(0, \frac{1}{\lambda}\mathbb{I})$ für ein $\lambda > 0$. Dann gilt:

$$F_x(\theta) := LL_x(\theta) - \lambda \cdot ||\theta||^2 + const.,$$

wobei $||\theta||$ die euklidische Norm des Vektors θ ist.

Berechnet man also den MAP-Schätzwert für θ, so maximiert man $F_x(\theta)$, d.h.

- maximiere die Likelihood,
- minimiere die Parameterwerte θ_i.

Im Vergleich zu MLE führt MAP mit normalverteilter a-priori Verteilung deswegen zu kleineren Werten für die Parameter.

Ausblick: Regularisierung im Machine Learning

Im Machine Learning schätzt man die Parameter einer Wahrscheinlichkeitsverteilung anhand von Daten (= Stichproben). Oft hat man dabei sehr viele Parameter (bis zu Millionen). Um sicherzustellen, dass die Vorhersagen stabil laufen, will man, dass diese Parameter keine zu extremen Werte annehmen - das nennt man im Machine Learning "Regularisierung".

Folgerung aus dem letzten Theorem: Eine Möglichkeit der Regularisierung ist, für die Parameterschätzung statt MLE MAP zu verwenden!