Zusammenfassung

In diesem Paper möchten wir eine kurze Einführung in die Bayes Theorie im Rahmen des Machine Learnings geben, welche den Anspruch hat für Leser mit nur geringen Kenntnissen der Statistik und Wahrscheinlichkeitstheorie lesbar zu sein. Wir nutzen die Bayes Theorie, um alternative numerische Problemstellungen zu "klassischen"Verfahren des Machine Learning herzuleiten. Exemplarisch führen wir aufbauend auf der linearen Regression in die Bayesianische Regression ein, und erhalten die regularisierten Verfahren "Ridge Regressionünd "LASSO". Weiterhin greifen wir die Klassifizierungstechnik der Support Vector Machine (SVM) auf und geben eine Herleitung der sogenannten "Least-Squares-SVM"(LS-SVM) an. Weiterführend werden Möglichkeiten zur Hyperparameterschätzung, sowie der Selektion von Kernen in diesem Modell erarbeitet. Diese Techniken münden abschließend gemeinsam in einen Algorithmus zur Lösung der LS-SVM bei mehreren konkurrierenden Kernen.

0.0.1 Einführung in die Bayes Theorie

Das Maschinelle Lernen nimmt in dem Zeitalter der Digitalen Revolution eine zunehmend zentrale Rolle ein. Schon jetzt basieren zahlreiche Produkte von global Playern wie Google, Facebook, aber auch Kalashnikov und Co. auf Techniken des Maschinellen Lernens. Zu diesen Techniken zählen Klassifizierungsmethoden wie die Support Vector Machine (SVM) und sogenannte Neural Networks. Eines ihrer charakteristischen Merkmale ist die Verwendung großer Datensätze (Big Data), was zu interessanten Fragestellungen im Bereich der Mathematik und Statistik, aber auch der Informatik führt. Wir möchten in diesem Paper eine mathematisch-statistische Perspektive auf gewisse Problemklassen einnehmen, um hieraus einen erweiterten Zugang zu "klassischen"numerischen Problemstellungen des Machine Learning zu erhalten.

Ziel ist es, bestimmte Sachverhalte mit Hilfe von Modellgleichungen abzubilden. Diese Modelle bestehen in der Regel aus Inputdaten $x \in \mathbb{R}^m$, einem Output $y \in \mathbb{R}^p$, sowie von Modellparametern, welche wir mit $\theta \in \mathbb{R}^k$ bezeichnen werden. Weiterhin werden wir Datensätze $D = (d_1, \cdots, d_n), d_i \in \mathbb{R}^l$ verwenden, um unsere Modelle zu trainieren.

Hierzu wählen wir den statistischen Ansatz der sogenannten Bayes Theorie. Diese Theorie verwendet Verteilungsannahmen an Parameter θ des Modells. Solche Annahmen ermöglichen eine umfangreiche wahrscheinlichkeitstheoretische Behandlung des Modells.

Zunächst einigen wir uns auf folgende Notation von Verteilungen und Wahrscheinlichkeiten:
Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein hinreichend großer Wahrscheinlichkeitsraum, denn wir im Folgenden immer im Hintergrund voraussetzen werden. Weiterhin sei $\theta : \Omega \to \mathbb{R}^k$ eine Zufallsvariable. Dann bezeichnen wir mit $\mathbb{P}(\theta) := \mathbb{P}^\theta = \mathbb{P}(\{\theta \in \cdot\})$ das Bildmaß der Zufallsvariable θ, also Ihre Verteilung. Wir möchten hier bemerken, dass es keinen Sinn macht nach der Wahrscheinlichkeit von θ zu fragen, sondern lediglich von der Wahrscheinlichkeit dass θ bestimmte Werte annimmt. Dies ist wichtig, um Missverständnisse zu vermeiden.

Wir möchten nun in die Grundbegriffe der Bayes Theorie einführen. Die zentralen Objekte der Bayes Theorie sind die A Priori Verteilung (Prior) und die A Posteriori Verteilung (Posterior), welche durch den Satz von Bayes in Verbindung stehen.

Definition 1 (A Priori Verteilung)
Sei der Modellparameter θ mit Werten im $\mathbb{R}^k$ eine Zufallsvariable.
Dann heißt seine Verteilung $\mathbb{P}(\theta)$ *A Priori Verteilung*, kurz *Prior*.

Der Prior als Wahrscheinlichkeitsverteilung muss vor dem weiteren Arbeiten mit einem Bayesianischen Modell durch eine Annahme festgelegt werden. Es ist für sich nicht klar welcher Prior für ein gegebenes Modell sinnvoll ist, was zur Gefahr widersprüchlicher oder hinderlicher Annahmen führt. Deshalb ist vor dem Platzieren eines Priors eine gründliche Überlegung über Eigenschaften der Parameter wichtig. Verschiedene Priors führen zu verschiedenen Modellen und somit zu verschiedenen Problemklassen, weshalb auch eine entgegengesetzte Betrachtung möglich ist, bei der bei gegebener Problemklasse der zugrunde liegende Prior gesucht ist. Wir wählen den anderen Ansatz, und konstruieren Probleme bei gegebenem Prior.

Bevor wir die A Posteriori Verteilung einführen, erinnern wir an den Satz von Bayes.

Theorem 2 (Satz von Bayes (Bayes 1763, Laplace 1812))
Sei $(\Omega, \mathcal{A}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, seien $A, B \in \mathcal{A}$ Ereignisse. Dann gilt für die bedingten Wahrscheinlichkeiten

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Falls $\mathbb{P}(B) = 0$, so definieren wir den rechten Ausdruck als 0.

Beweis. Mit Definition einer bedingten Wahrscheinlichkeit erhält man

$$\frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\frac{\mathbb{P}(B\cap A)}{\mathbb{P}(A)}\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B\cap A)}{\mathbb{P}(B)} = \mathbb{P}(A|B).$$

□

Gegeben eines Priors möchte man nun die statistischen Daten eines Modells in die Information über die Parameter miteinfliessen lassen. Die hierzu folgende Definition ist in Anlehnung an den Satz von Bayes motiviert.

Definition 3 (A Posteriori Verteilung)
Sei θ ein Modellparameter mit A Priori Verteilung $\mathbb{P}(\theta)$, $D = (d_1, \cdots, d_n)$ ein Datenvektor. Dann definieren wir die *A Posteriori Verteilung*, kurz *Posterior*, durch

$$\mathbb{P}(\theta|D) := \frac{\mathbb{P}(D|\theta)\mathbb{P}(\theta)}{\mathbb{P}(D)},$$

wobei $\mathbb{P}(D) = \int \mathbb{P}(D|\theta) d\mathbb{P}(\theta)$.

In dem Posterior sind die kombinierte statistische Information der Annahmen, sowie der beobachteten Daten, über θ enthalten. Der Posterior besteht aus der Likelihood $\mathbb{P}(D|\theta)$, dem Prior $\mathbb{P}(\theta)$, sowie einem Normalisierungsfaktor $\mathbb{P}(D)$.

Der Prozess der Berechnung des Posteriors wird oft auch *Training* genannt, in diesem Kontext nennt man D oft auch *Trainingsdaten*. Die exakte Berechnung des Posteriors gestaltet sich bei großen Datenmengen D als schwierig, besonders wenn sogenannte latente Variablen im Modell auftreten. Latente Variablen Z sind Variablen, die nicht beobachtet werden können, und bei denen somit keine Daten zur Verfügung stehen. Diese treten häufig im Kontext von Bayesianischen Netzwerken, Modellen zur Textkategorisierung, rekurrenten Neuronalen Netzwerken und Sprach- sowie Texterkennung auf. Für einführende Beispiele

von Modellen mit latenten Variablen im Machine Learning verweisen wir beispielhaft auf den Vortrag [**?**].

Es bestehen Möglichkeiten zur approximativen Bestimmung des Posteriors. Als Beispiel einer statistischen Methode zur approximativen Berechnung seien hier die *Markov-Chain-Monte-Carlo-Verfahren* genannt. Weiterhin gibt es analytische Approximationen, welche auf große Klassen von Verteilungen angewandt werden können. Beispielhafte Techniken sind die *Expectation-Maximization-Algorithmen* und *Variational-Bayes-Methoden*. Für eine Einführung in erstere, siehe [**?**], für eine ausführliche Behandlung des letzteren im Rahmen des Machine Learning siehe [**?**].

Wir nutzen den nun eingeführten Posterior um eine Bayesianische Art von Parameterschätzer zu definieren, den sogenannten Maximum-A-Posteriori-Schätzer. Zuvor wiederholen wir Begriff des Maximum-Likelihood-Schätzers.

Definition 4
Maximum-Likelihood-Schätzer Sei $D = (d_1, \cdots, d_n)$ ein Datensatz, θ ein $\mathbb{R}^k$-wertiger Parameter. Dann heißt der Schätzer

$$\hat{\theta}_{MLE} := \arg\max_{\theta} \mathbb{P}(D|\theta)$$

Maximum-Likelihood-Schätzer, kurz *MLE-Schätzer*.

Es sei hier erwähnt, dass ein tiefer Zusammenhang der Parameterschätzung und Optimierungsmethoden auf Mannigfaligketien besteht. Betrachtet man den Raum aller Parameter $\theta \in \Theta$ und versieht diesen mit der Fisher-Informations-Metrik, so erhält man eine riemannsche Mannigfaltigkeit, dessen Punkte als die zu den Parametern gehörigen Verteilungen interpretiert werden können. Y. Ollivier et. al. (2017) führen auf diese Weise ein Schema mit Black-Box-Ansatz zur Konstruktion von Optimierungsmethoden mit Hilfe des zur Fisher-Informations-Metrik gehörenden natürlichen Gradienten ein. Leser mit Interesse an Differentialgeometrie und numerischer Optimierung verweisen wir auf [**?**]. Für eine kurze Einführung in das Gebiet der Informationsgeometrie verweisen wir auf [**?**].

Nun führen wir in Analogie zu dem MLE-Schätzer den Bayesianischen Maximum-A-Posteriori-Schätzer ein.

Definition 5 (Maximum-A-Posteriori-Schätzer)
Sei $D = (d_1, \cdots, d_n)$ ein Datenvektor, θ ein $\mathbb{R}^k$-wertiger Parameter mit A Priori Verteilung $\mathbb{P}(\theta)$. Dann heißt der Schätzer

$$\hat{\theta}_{MAP} := \arg\max_{\theta} \mathbb{P}(\theta|D)$$

Maximum-A-Posteriori-Schätzer, kurz *MAP-Schätzer*. Hierbei ist $\mathbb{P}(\theta|D)$ der Posterior von θ gegeben D.

Der MAP-Schätzer ist ein statistischer Schätzer, der sich von dem MLE-Schätzer dadurch unterscheidet, dass er sowohl A Priori Annahmen an die Verteilung, als auch Daten bei der Parameterschätzung berücksichtigt. Diesen Zusammenhang erkennt man an folgender Proposition:

Proposition 6 (Darstellung des MAP-Schätzers)

Es gelten folgende Identitäten des MAP-Schätzers:

$$\begin{aligned}\hat{\theta}_{MAP} :&= \arg\max_{\theta} \mathbb{P}(\theta|D) \\ &= \arg\max_{\theta\in\Theta} \frac{\mathbb{P}(x|\theta)\mathbb{P}(\theta)}{\int \mathbb{P}(x|\theta)d\mathbb{P}(\theta)} \\ &= \arg\max_{\theta\in\Theta} \mathbb{P}(x|\theta)\mathbb{P}(\theta) \\ &= \arg\max_{\theta\in\Theta} \log \mathbb{P}(x|\theta) + \log \mathbb{P}(\theta)\end{aligned}$$

Beweis. Definition des MAP-Schätzers und des Posteriors, sowie die Tatsache, dass $\int \mathbb{P}(x|\theta)d\mathbb{P}(\theta) \in \mathbb{R}$ als Skalar nicht von θ abhängt und log monoton wachsend ist. □

Besonders anhand der letzten Identität erkennt man, dass der MAP-Schätzer die Information der Daten D im log-Likelihood-Anteil mit der A Priori-Information des log-transformierten Priors kombiniert. Diese Darstellungen werden wir im weiteren verwenden um Optimierungsprobleme. Im Wesentlichen wird es bei allen Herleitungen in den folgenden Kapiteln darum gehen, durch das Aufstellen des MAP-Schätzers in verschiedenen Situationen regularisierte oder gänzlich neue Problemstellungen und Verfahren im Machine Learning herzuleiten. Da wir nun die Grundlagen eingeführt haben, fahren wir fort mit unserem ersten nichttrivialen Anwendungsbeispiel, der Bayesianischen Regression.