Machine Learning Compendium

Machine Learning University Society at Karlsruhe Institute of Technology (KIT)

Machine Learning Compendium

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Morbi commodo, ipsum sed pharetra gravida, orci magna rhoncus neque, id pulvinar odio lorem non turpis. Nullam sit amet enim. Suspendisse id velit vitae ligula volutpat condimentum. Aliquam erat volutpat. Sed quis velit. Nulla facilisi. Nulla libero. Vivamus pharetra posuere sapien. Nam consectetuer. Sed aliquam, nunc eget euismod ullamcorper, lectus nunc ullamcorper orci, fermentum bibendum enim nibh eget ipsum. Donec porttitor ligula eu dolor. Maecenas vitae nulla consequat libero cursus venenatis. Nam magna enim, accumsan eu, blandit sed, blandit a, eros.

Contents

Authors

The following authors contributed to Chapter 1: Sandro Braun

The following authors contributed to Chapter 2: Leander Kurscheidt

1

Reinforcement Learning

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. wikipedia

1.1 Bellman Equations

TODO backup diagramms

1.1.1 State Value Function

$$v_\pi(s) = \sum_{a \in A} \pi(s|a) Q_\pi(s,a) \tag{1.1}$$

1.1.2 Action Value Function

$$Q_\pi(s,a) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s') \tag{1.2}$$

1.1.3 State Value Function recursive

$$v_\pi(s) = \sum_{a \in A} \pi(s|a)(r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')) \tag{1.3}$$

1.1.4 Action Value Function recursive

$$Q_\pi(s,a) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a \in A} \pi(a'|s') Q_\pi(s',a') \tag{1.4}$$

1.1.5 Optimal State Value Function

$$v_*(s) = \max_a Q_*(s, a) \tag{1.5}$$

1.1.6 Optimal Action State Value Function

$$Q_*(s, a) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s') \tag{1.6}$$

1.1.7 Optimal State Value Function recursive

$$v_*(s) = \max_a r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s') \tag{1.7}$$

1.1.8 Optimal Action State Value Function recursive

$$Q_*(a, s) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \max_{a'} Q_*(s', a') \tag{1.8}$$

1.2 Advantage Function

TODO

1.3 Policy, Policy Gradient

1.3.1 Policy: Distribution over actions given states

$$\pi_\theta(a|s) = P(a|s) \tag{1.9}$$

1.3.2 Policy Gradient

$$\nabla_\theta \pi_\theta(s|a) = \pi_\theta(s|a) \nabla_\theta \log \pi_\theta(s|a) \tag{1.10}$$

Note: this is valid for all probability distributions (the policy is a distribution over actions given states). The gradient term on the right hand side is called score function. The derivation basically uses the "log-trick".

2

Statistics

Statistics is a branch of mathematics dealing with the collection, organization, analysis, interpretation and presentation of data.

wikipedia

2.1 Probability

TODO: Probability (general + simple), CDF, Variance, etc.

2.1.1 L_p-Space for Random-Variables

The L_p-Norm for Random-Variables X, where $\mathbb{E}|X|^p < \infty$, is defined through:

$$||X||_p := (\mathbb{E}[|X|^p])^{\frac{1}{p}}$$

lecture

2.1.2 Jensens-Inequality for Random Variables

If ϕ is a konvex function and X a Random-Variable, then

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X)$$

wikipedia

2.1.3 Fisher-Information

For the parametric family $\mathcal{P} \in \{\mathcal{P}_\theta | \theta \in \Theta_L\}$...TODO

If we assume $p_\theta(x, y) = p_\theta(y|x)p(x)$, the Fisher-Information Matrix $\mathcal{I}(\theta)$ becomes:

$$\mathcal{I}(\theta) = \mathbb{E}_{(X,Y)\sim P_\theta}[\nabla_\theta \log p_\theta(Y|X) \otimes \nabla_\theta \log p_\theta(Y|X)],$$

where $\otimes$ is the inner-product.

paper

2.2 Distributions

In this section, X denotes a Random Vairable and f the density-function.

TODO: More common distributions

2.2.1 Normal Distribution

If $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$ and $\sigma^2 > 0 \in \mathbb{R}$, then:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mathbb{E}X = \mu$$

$$Var[X] = \sigma^2$$

wikipedia

2.2.2 Normal Distribution (Multivariate)

If $X \sim \mathcal{N}(\mu, \Sigma)$ for $\mu \in \mathbb{R}^k$ and $\Sigma \in \mathbb{R}^{k \times k}$ with Σ being positve semi-definite, then:

$$f(x) = \det(2\pi\mathbf{\Sigma})^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$$\mathbb{E}X = \mu$$

$$Var[X] = \Sigma$$

wikipedia

2.2.3 Empirical Distribution

For any observation $X' = (x_1', \cdots, x_n')$, the empirical distribution is defined as:

$$f(x) = \hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i), \text{ where } \delta \text{ is the dirac-delta function}$$

$$\mathbb{E}X = \hat{\mathbb{E}}X = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Var[X] = V\hat{a}r[X] = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mathbb{E}}X)^2$$

lecture

2.3 Estimation

TODO: ML, Score-Function, biased/unbiased, CramérRao bound, confidence-interval

2.4 Divergences

Conventions for this section: P and Q are probability measures over a set X, and P is absolutely continuous with respect to Q. S is a space of all probability distributions with common support.

2.4.1 Divergence

A divergence on S is a function $D : S \times S \to R$ satisfying

1. $D(p||q) \geq 0 \forall p, q \in S$,
2. $D(p||q) = 0 \Leftrightarrow p = q$

A divergence is a "sense" of distance between two probability distributions. It's not a metric, but a pre-metric.

wikipedia

2.4.2 f-Divergence

1. Generalization of whole family of divergences

2. For a convex function f such that $f(1) = 0$, the f-divergence of P from Q is defined as:
   $D_f(P \parallel Q) \equiv \int_\Omega f\left(\frac{dP}{dQ}\right) dQ$

3. wikipedia

2.4.3 KL-Divergence

1. The KullbackLeibler divergence from Q to P is defined as $D_{\mathrm{KL}}(P\|Q) = \int_X \log \frac{dP}{dQ}\, dP = D_{t \log t}$.

2. maxmizing likelihood is equivalent to minimizing $D_{KL}(P(.|\theta^*) \parallel P(.|\theta))$ (the foreward-KL Divergence), where $P(.|\theta^*)$ is the true distribution and $P(.|\theta)$ is our estimate.

3. wikipedia

4. TODO: Fisher-Matrix infitesimal relationship

2.4.4 JensenShannon divergence

The JensenShannon divergence from Q to P is defined as

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$$

, where $M = \frac{1}{2}(P + Q)$

wikipedia

2.4.5 TODO: Wasserstein & Wasserstein Dual

2.5 Information Geometry

Information Geometry defines a Riemannian Manifold over probability distributions for statistical models.

2.5.1 Fisher-Rao Metric

For the parametric family $\mathcal{P} \in \{\mathcal{P}_\theta | \theta \in \Theta_L\}$ and every $\alpha, \beta \in \mathbb{R}^d$ with their tangent-vectors $\bar{\alpha} = dp_{\theta+t\alpha}/dt|_{t=0}$ and $\bar{\beta} = dp_{\theta+t\beta}/dt|_{t=0}$, we define the inner local product as follows:

$$
\begin{aligned}
< \bar{\alpha}, \bar{\beta} > &:= \int_M \frac{\bar{\alpha}}{p_\theta} \frac{\bar{\beta}}{p_\theta} p_\theta \\
&= < \alpha, \mathcal{I}(\theta)\beta > ,
\end{aligned}
$$

where $\mathcal{I}(\theta)$ is the Fisher-Information Matrix.

2.5.2 Natural Gradient

The natural gradient is the gradient descent induced by the Fisher-Rao geometry of $\{\mathcal{P}_\theta\}$.

paper