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

Statistics

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

wikipedia

1.1 Probability

TODO: Probability (general + simple), CDF, Variance, Markov-Property, stationarity, etc.

1.1.1 Statistic

A statistic is a function of a sample where the function itself is independent of the sample's distribution; that is, the function can be stated before realization of the data. The term statistic is used both for the function and for the value of the function on a given sample. wikipedia

1.1.2 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

1.1.3 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

1.1.4 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

1.2 Distributions

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

TODO: More common distributions

1.2.1 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

1.2.2 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

1.2.3 Empirical Distribution

For any observation $X' = (x'_1, \cdots, x'_n)$, the empirical distribution is defined as:

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

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

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

lecture

1.3 Estimation

TODO: biased/unbiased, consistent, sufficient, CramérRao bound, confidence-interval

1.3.1 Estimator

Examples:

1. *point estimation:*
   The application of a point estimator (a statistic) to the data to obtain a point estimate. In Machine-Learning, estimating the parameters of neural networks is usually done via (a multidimensional) point-estimation. wikipedia

2. *Interval estimation:*
   interval estimation is the use of sample data to calculate an interval of plausible values of an unknown population parameter.wikipedia

3. *clustering:*
   Grouping data into sets of similiar objects.

4. *classification:*
   Assigning Categories to data-objects.

1.3.2 Score Function

1. indicates how sensitive a likelihood function $\mathcal{L}(\theta; X)$ is to its parameter θ.

2. it is defined as:

$$u_\theta(x) = \frac{\partial}{\partial\theta} ln\mathcal{L}(\theta \mid x)$$

   where $\mathcal{L}$ is a likelihood-function.

wikipedia

1.3.3 Likelihood

1. A likelihood function (often simply the likelihood) is a function of the parameters of a statistical model, given specific observed data

2. Common definitions:

   (a) *parameterized model*
   Given a parameterized family of probability density functions (or probability mass functions in the case of

discrete distributions) $x \mapsto f(x \mid \theta)$, where θ is the parameter, the likelihood function is

$$\theta \mapsto f(x \mid \theta)$$

, written

$$\mathcal{L}(\theta \mid x) = f(x \mid \theta)$$

where x is the observed outcome of an experiment.

(b) *In general*
The likelihood function is this density interpreted as a function of the parameter (possibly a vector), not of the possible outcomes. This provides a likelihood function for any probability model with all distributions, whether discrete, absolutely continuous, a mixture or something else.

3. *Log-likelihood:*
It's usually convenient to work with the log-likelihood, especially if multiple, indipendent random variables are involed.

wikipedia

1.3.4 Maximum-Likelihood Estimator

1. Maximum likelihood estimation (MLE) attempts to find the parameter values that maximize the likelihood function, given the observations.

2. *frequentist inference:* MLE is one of several methods to get estimates of parameters without using prior distributions.

Some properties:

1. *Consistency:* the sequence of MLEs converges in probability to the value being estimated.
2. *Efficiency:* it achieves the CramérRao lower bound when the sample size tends to infinity. This means that no consistent estimator has lower asymptotic mean squared error than the MLE (or other estimators attaining this bound).
3. *Second-order efficiency* after correction for bias.

wikipedia

1.4 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.

1.4.1 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

1.4.2 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

1.4.3 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

1.4.4 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

1.4.5 Wasserstein Metric

1. Also often called Earth-Mover's distance (EDM)
2. It differs from the usual KL-Divergence in that it's based on optimal-transport and not on local probability differences

Let (M, d) be a metric space with every probability measure on M being a Radon measure (a so-called Radon space). For $p \geq 1$, let $P_p(M)$ denote the collection of all probability measures μ on M with finite p^{th} moment for some x_0 in M.

Wasserstein Metric - Primal

The p^{th} Wasserstein distance between two probability measures μ and ν in $P_p(M)$ is defined as

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}$$

, where $\Gamma(\mu, \nu)$ denotes the collection of all measures on $M \times M$ with marginals μ and ν. ($\Gamma(\mu, \nu)$ is also called the set of all couplings of μ and ν .)

Wasserstein Metric - Dual

The following dual representation of W_1 is a special case of the duality theorem of Kantorovich and Rubinstein (1958): when μ and ν have bounded support,

$$W_1(\mu,\nu) = \sup\left\{ \int_M f(x)\,\mathrm{d}(\mu-\nu)(x) \middle| \text{continuous } f : M \to \mathbb{R}, \mathrm{Lip}(f) \leq 1 \right\}$$

, where $Lip(f)$ denotes the minimal Lipschitz constant for f.

If the metric d is bounded by some constant C, then $2W_1(\mu,\nu) \leq C\rho(\mu,\nu)$, and so convergence in the Radon metric (identical to total variation convergence when M is a Polish space) implies convergence in the Wasserstein metric, but not vice versa.

This article uses material from the Wikipedia article Wasserstein metric which is released under the Creative Commons Attribution-Share-Alike License 3.0.

Good introductionary blog-post: wasserstein.

1.5 Information Geometry

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

1.5.1 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.

1.5.2 Natural Gradient

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

paper

2

Neural Network

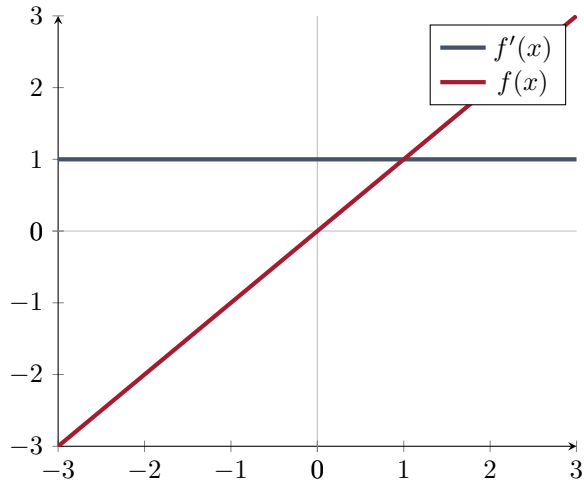
Artificial neural networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. wikipedia TODO: lot of stuff

2.1 Architecture

TODO: Perceptron, Fully-Connected, Convolutional, Autoencoder, LSTM etc.

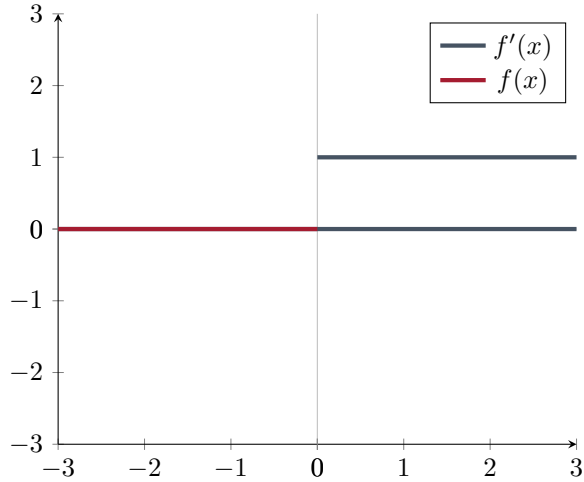
2.1.1 Identity

definition	derivative
$f(x) = x$	$f'(x) = 1$



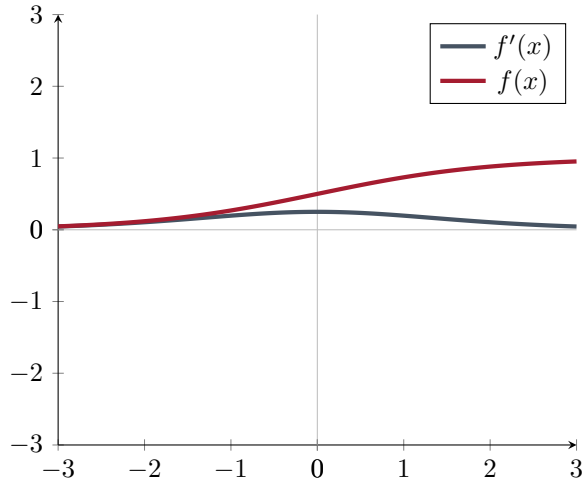
2.1.2 Heav-Step function

definition	derivative
$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & else \end{cases}$	$f'(x) = 0$



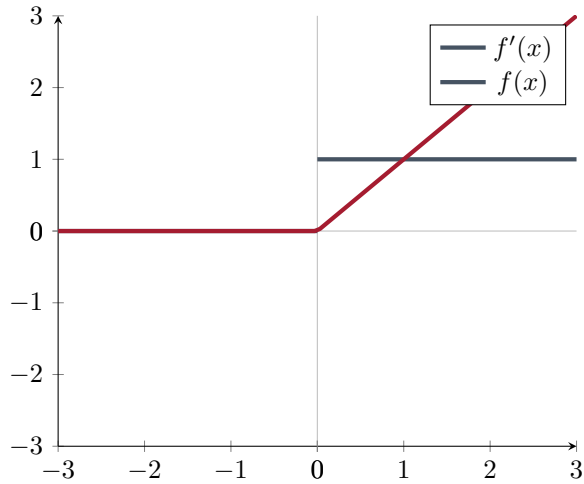
2.1.3 Logistic

definition	derivative
$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = \frac{e^{-x}}{(e^{-x}+1)^2} = f(x)(1 - f(x))$



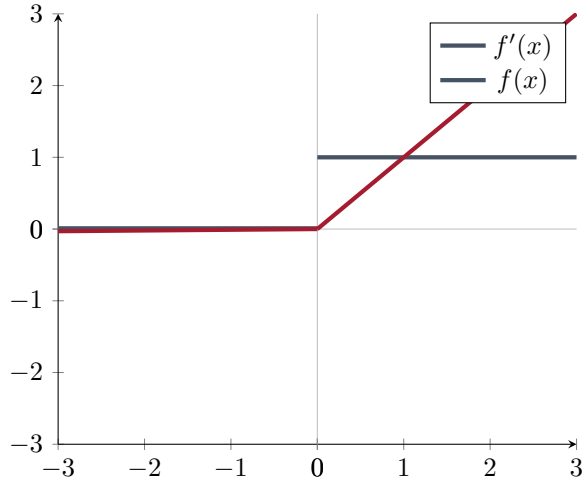
2.1.4 ReLu

definition	derivative
$f(x) = max(0, x)$	$f'(x) = \begin{cases} 1 & x \geq 0 \\ 0 & else \end{cases}$



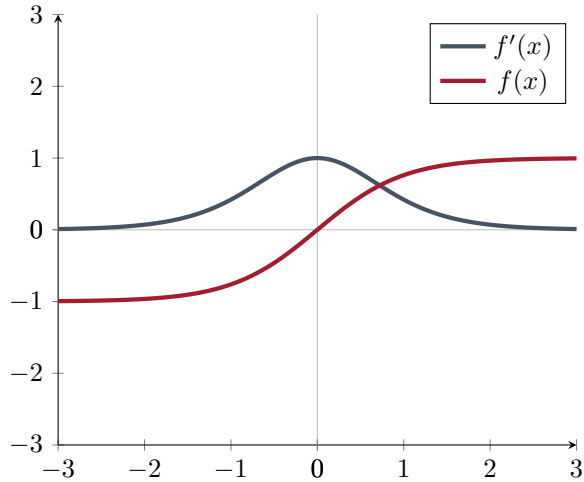
2.1.5 Leaky ReLu

definition	derivative
$f(x) = \begin{cases} x & x \geq 0 \\ 0.01x & else \end{cases}$	$f'(x) = \begin{cases} 1 & x \geq 0 \\ 0.01 & else \end{cases}$



2.1.6 Tanh

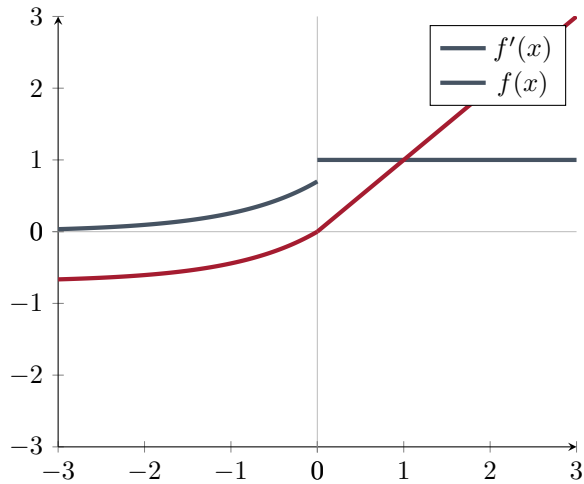
definition	derivative
$f(x) = tanh(x)$	$f'(x) = 1 - tanh(x)^2$



2.1.7 ELU

the examples are plottet for $\alpha = 0.7$.

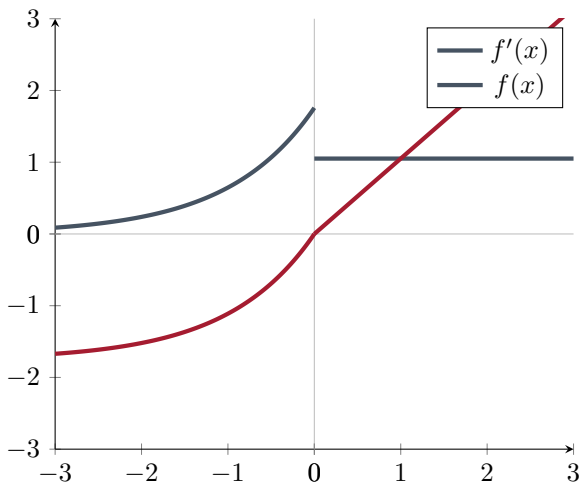
definition	derivative
$f(x) = \begin{cases} \alpha(e^x - 1) & x \geq 0 \\ x & else \end{cases}$	$f'(x) = \begin{cases} 1 & x \geq 0 \\ \alpha * e^x = f(x) + \alpha & else \end{cases}$



2.1.8 SELU

Scaled Exponential Linear Unit (SELU) is the Exponetial Linear Unit (ELU) activation function, where λ and α have been fixed to 1.0507 and 1.67326. Neural Networks using the SELU form Self-Normalizing Neural Networks. See the paper for more information.

definition	derivative
$f(x) = \lambda \begin{cases} \alpha(e^x - 1) & x \geq 0 \\ x & else \end{cases}$	$f'(x) = \lambda \begin{cases} 1 & x \geq 0 \\ \alpha * e^x = f(x) + \alpha & else \end{cases}$



3

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

3.1 Bellman Equations

TODO backup diagramms

3.1.1 State Value Function

$$v_\pi(s) = \sum_{a \in A} \pi(s|a) Q_\pi(s,a) \tag{3.1}$$

3.1.2 Action Value Function

$$Q_\pi(s,a) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s') \tag{3.2}$$

3.1.3 State Value Function recursive

$$v_\pi(s) = \sum_{a \in A} \pi(s|a)(r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')) \tag{3.3}$$

3.1.4 Action Value Function recursive

$$Q_\pi(s,a) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a \in A} \pi(a'|s') Q_\pi(s',a') \tag{3.4}$$

3.1.5 Optimal State Value Function

$$v_*(s) = \max_a Q_*(s, a) \tag{3.5}$$

3.1.6 Optimal Action State Value Function

$$Q_*(s, a) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s') \tag{3.6}$$

3.1.7 Optimal State Value Function recursive

$$v_*(s) = \max_a r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s') \tag{3.7}$$

3.1.8 Optimal Action State Value Function recursive

$$Q_*(a, s) = r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \max_{a'} Q_*(s', a') \tag{3.8}$$

3.2 Advantage Function

TODO

3.3 Policy, Policy Gradient

3.3.1 Policy: Distribution over actions given states

$$\pi_\theta(a|s) = P(a|s) \tag{3.9}$$

3.3.2 Policy Gradient

$$\nabla_\theta \pi_\theta(s|a) = \pi_\theta(s|a) \nabla_\theta \log \pi_\theta(s|a) \tag{3.10}$$

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".