B. Discrete Valued RV. PMF

$\underline{X} \in \{\underline{X}_1, \dots \underline{X}_c\} \sim P(\underline{x})$: True PMF

$Q(\underline{x})$: Approx for $P(\underline{x})$

$$D_{KL}(P\|Q) = E_{\underline{X}\sim p}\left[\ln \frac{p(\underline{X})}{q(\underline{X})}\right]$$

$$= \sum_{i=1}^{N} p(x_i) \ln \frac{p(\underline{x}_i)}{q(\underline{x}_i)}$$

Properties of KLD:

P1. Non - negative : $D_{KL}(p\|q) \geq 0 \ \forall p, q$

P2: Equality:

$D_{KL}(p\|q) = 0$ iff $p(\underline{x}) = q(\underline{x}) \ \forall \ \underline{x}$

Proof of "sufficient": $\ln \frac{p(x)}{q(x)} = \ln 1 = 0 \ \forall x$

P1 + P2:

KLD is a suitable metric for approximating p by q.

P3. Assymmetric :

$D_{KL}(p\|q) \neq D_{KL}(q\|p)$

$= E_{\underline{X}\sim p} \ln\left(\frac{p(\underline{x})}{q(\underline{x})}\right) \qquad = E_{x\sim q} \ln\left(\frac{q(x)}{p(\underline{x})}\right)$

Hence it is a divergence, not distance

forward KLD

backward KLD

D_{KL} is <u>NOT</u> a true distance measure, with

$D(\underline{x}, \underline{y}) \neq D(\underline{y}, \underline{x})$.

We mainly use forward DKL.

To minimize DKL, always increase the denominator in the ln( ) term.

$$\left\{ \ln\left(\frac{p(x)}{q(x)\uparrow}\right)\downarrow \right\}\downarrow$$

$\boxed{\text{Exercise}}$

- <u>KLD between Gaussian and Laplace distribution</u>

$p(x) \sim N(0, \sigma^2)$, $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}x^2}$

$q(x) \sim Laplace(0, b)$, $q(x) = \frac{1}{2b} e^{-\frac{|x|}{b}}$

We cant get $D_{KL} = 0$ as both are different

Choose b such as to best approximate $p(x)$ by $q(x)$.

$$\frac{p(x)}{q(x)} = \frac{2b}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{|x|}{b}\right)$$

$$\ln\left(\frac{p(x)}{q(x)}\right) = \ln\left(\sqrt{\frac{2}{\pi}}\frac{b}{\sigma}\right) - \frac{x^2}{2\sigma^2} + \frac{|x|}{b}$$

$$D_{KL}(p\|q) = E_{x\sim p} \ln\left(\frac{p(x)}{q(x)}\right) = \ln\left(\sqrt{\frac{2}{\pi}}\frac{b}{\sigma}\right) + E_{x\sim p}\left(-\frac{x^2}{2\sigma^2} + \frac{|x|}{b}\right)$$

$$E_{x\sim p}(|X|) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \cdot |x|\, dx$$

$\vdots$

$$= \sqrt{\frac{2}{\pi}}\,\sigma$$

$E_{x\sim p}(X^2) = \sigma^2$

If $\alpha = \frac{\sigma}{b}$, Then $D_{KL}(p\|q) = \ln\left(\sqrt{\frac{2}{\pi}}\right) - \ln\alpha - \frac{1}{2} + \sqrt{\frac{2}{\pi}}\alpha$

Now we need to min α.

$$\frac{\partial D_{KL}}{\partial \alpha} = -\frac{1}{\alpha} + \sqrt{\frac{2}{\pi}} \stackrel{!}{=} 0$$

$$\Rightarrow \alpha = \sqrt{\frac{\pi}{2}}$$

$$\approx b \approx 0.8 \, \sigma$$

$$D_{KL}(p \| q) \Big|_{\alpha = \sqrt{\frac{\pi}{2}}} = -\ln \frac{\pi}{2} + \frac{1}{2}$$

$$\approx 0.048$$

Deep Learning

Assignments - Online - Mandatory for Lab.

"A recipe ... " - VVI - download and read.

1. Introduction

Slide 1.0

1.1 What is machine learning?

Slide 1.2 - 1.3

Signal to signal → noised audio to denoised audio } Regression

" to parameter → auto driving eg range of radar } (Continuous)

" to class → classification → finite no of categories

→ see radar workings papers → from signal $x(n)$ to distⁿ of target
Sync using correlation.

Slide 1.9 — Overfitting is to be avoided so divide training ds to training + validation.

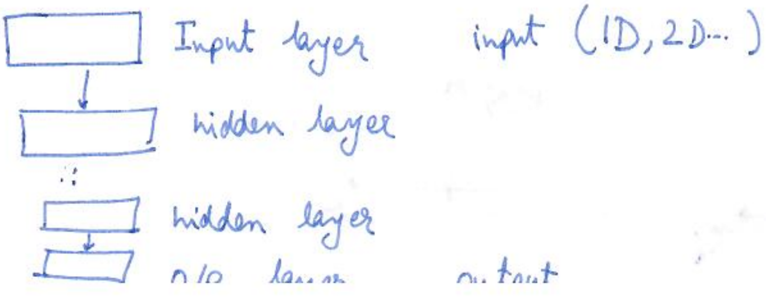
1.10 → mainly we will focus on supervised learning ds.
unsupervised → DPR

1.11.1

What is deep learning?

1.12 → Conventional ML is in DPR

1.13.1 → What is a neural network? (NN)



14.04.2022

Ch 2 Tools for Deep Learning

2.1 Software

Scipy, numpy, Tensorflow

2.2 Hardware

Use Google colab for assignments

2.3 Datasets

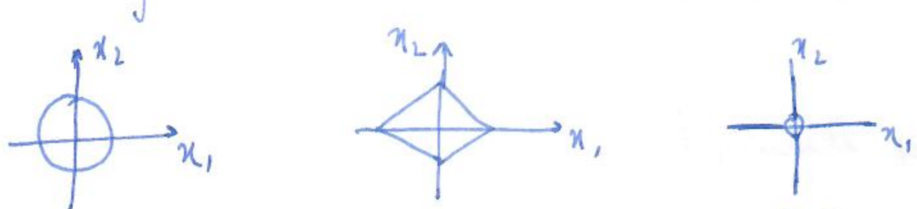
Q Cityscaps ds → why ~~cant~~ wasn't self-supervised learning not performing well ?

3. ML basics

3.1 Linear algebra

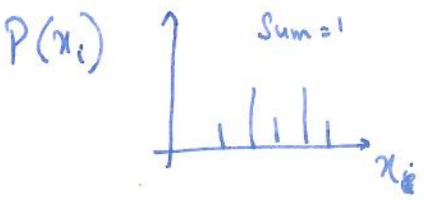
See AM for more details

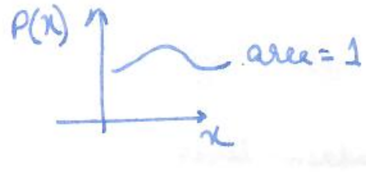
Mainly we use 1-norm in ML



$p=2: x_1^2+x_2^2=1$ $\quad$ $p=1: |x_1|+|x_2|=1$ $\quad$ $p=0:$ one of $x_1, x_2 \neq 0$

3.2 Random variables and probability distribution.





3.7.1

PDF for discrete valued RVs:

$P(\underline{x}) = \sum_i P_i \delta(\underline{x} - \underline{x}_i)$, $\delta(\underline{x})$: dirac f.n

Cumulative distribution f.n (CDF)

$$F(\underline{x}) = P(\underline{X} \leq \underline{x}) = \int_{-\infty}^{\underline{x}} p(\underline{z}) d\underline{z}$$

$$P(\underline{x}) = \frac{\partial^d F(\underline{x})}{\partial x_1 \dots \partial x_d}$$

3.9.1 ~~ex~~

Special case $d = 1$

$\underline{X} \rightarrow X \in \mathbb{R}$

$\underline{\mu} \rightarrow \mu \in \mathbb{R}$

$\underline{\underline{C}} \rightarrow$ variance of $\underline{X} = \mathrm{Var}(X) = \sigma^2$

$$= E[(X-\mu)^2] = E[X^2] - \mu^2$$

$\sigma = \sqrt{\mathrm{Var}(X)} \geq 0$: standard ~~varia~~ deviation of X

for any f.n $g(\underline{X})$ of $\underline{X}$:

$$E[g(x)] = \int g(\underline{x}) p(\underline{x}) d\underline{x} \stackrel{d.v}{=} \sum_i g(\underline{x}_i) P(\underline{x}_i)$$

3.12

Multinoulli dist^n. is an ~~as~~ extended case of Bernoulli dist^n

21.4.22

3.14 One-hot coding

1 - hot , 0 - cold

One-hot coding - used only in classification

3.15.1 Reformulation of categorical distribution by one-hot coding:

$\underline{x} = [x_i] \in \{\underline{e}_1, \dots \underline{e}_c\}$, all $x_i = 0$ except for one value at i^{th} position as 1.

$$\Rightarrow \text{PMF} \quad P(\underline{X} = \underline{x}) = P(\underline{X})$$

$$= \begin{cases} p_1 & \text{if } \underline{X} = \underline{e}_1 \quad \text{or } x_1 = 1 \\ \vdots & \quad \vdots \qquad\qquad \vdots \\ p_c & \text{if } \underline{X} = \underline{e}_c \quad \text{or } x_c = 1 \end{cases}$$

$$= p_1^{x_1} \cdots\cdots p_c^{x_c}$$

$$= \prod_{i=1}^{c} p_i^{x_i}$$

$$\ln P(\underline{X}) = \sum_{i=1}^{c} x_i \log p_i = [x_1, \dots x_c] \cdot \begin{bmatrix} \ln p_1 \\ \vdots \\ \ln p_c \end{bmatrix}$$

$$= \underline{x}^T \ln \underline{p}$$

element wise

3.2.2 Multiple RV

Product rule $p(\underline{x}, \underline{y}) = p(\underline{x} | \underline{y})\, p(\underline{y}) = p(\underline{y} | \underline{x})\, p(\underline{x})$

Bayes rule $p(\underline{y} | \underline{x}) = p(\underline{x} | \underline{y}) \cdot \dfrac{p(\underline{y})}{p(\underline{x})}$

3.17 Chain rule of Probability

$$p(\underline{x}_1, \dots \underline{x}_N) = p(\underline{x}_1 | \underbrace{\underline{x}_2 \dots \underline{x}_N}_{\underline{y} \leftarrow \text{it can be assumed}})\, p(\underline{x}_2 \dots \underline{x}_N)$$

Two RV $\underline{X}$ and $\underline{Y}$ independent if

$$p(\underline{x}, \underline{y}) = p(\underline{x})\, p(\underline{y})$$

$$\iff p(\underline{x} | \underline{y}) = p(\underline{x}) \quad , \quad p(\underline{y} | \underline{x}) = p(\underline{y})$$

$\underline{X}_1 \dots \underline{X}_N$ are independent and identically distrubed (iid)

$$* \quad p(\underline{x}_1 \dots \underline{x}_N) = \prod_{i=1}^{N} p_i(\underline{x}_i) \qquad \underline{X}_i \sim p_i(\underline{x}_i)$$

$$* \quad p_i(x_i) = p(x_i) \;\forall i$$

$$\Rightarrow p(\underline{x}_1 \dots x_N) = \prod_{i=1}^{N} p(\underline{x}_i)$$

3.2.3 Kernel-based density estimation

PDF $p(\underline{x})$ of $\underline{X} \in \mathbb{R}^d$ unknown. Only i.i.d samples

$\underline{x}(n)$ $. 1 \leq i \leq N$ available

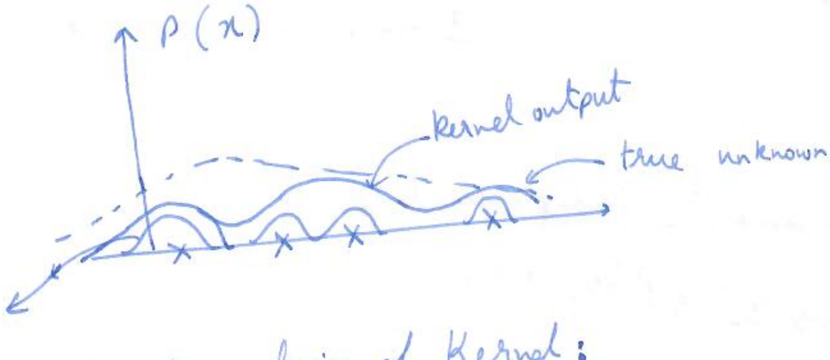
Kernel-based estimate of $p(\underline{x})$ from $\underline{x}(n)$:

Kernel $k(\underline{x})$, like a PDF

*) $k(\underline{x}) \geq 0 \quad \forall \underline{x}$

*) $\int k(\underline{x})\, d\underline{x} = 1$

$$\hat{p}(\underline{x}) = \frac{1}{N} \sum_{n=1}^{N} k(\underline{x} - \underline{x}(n))$$



popular choice of Kernel:

Gaussian kernel - smooth

*) Standard $N(0, \underline{\underline{I}})$: $k(\underline{x}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|\underline{x}\|^2}$,

fixed width

*) $N(\underline{0}, \sigma^2 \underline{\underline{I}})$: $k(\underline{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{1}{2\sigma^2}\|x^2\|}$,

variable ~~lena~~ width

Dirac kernel

$k(\underline{x}) = \delta(\underline{x})$

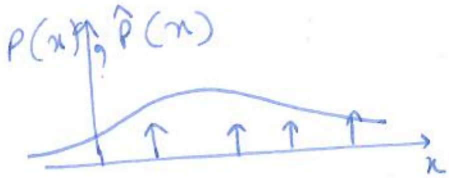
*) $\delta(\underline{x}) = \begin{cases} 0 & \underline{x} \neq 0 \\ \infty & \underline{x} = \underline{0} \end{cases}$

*) $\int \delta(\underline{x})\, d\underline{x} = 1$

*) Sampling property $\quad \int \delta(x - x_0) f(\underline{x})\, dx = f(x_0)$

Emperical distribution

$$\hat{p}(\underline{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\underline{x} - \underline{x}(n))$$



3.19.1

3.3 Kullback - Liebler divergence and cross - entropy

Dissimilarity measure b/w two distribution - KL divergence

A. Continuous - valued RV : PDF

$\underline{X} \sim p(\underline{x})$: true distribution of $\underline{X}$

$g(\underline{x})$: approximation for $p(\underline{x})$ by a DNN

KLD between p and g :

$$D_{KL}(p \| g) = \underbrace{\int p(\underline{x}) \ln \underbrace{\frac{p(\underline{x})}{g(\underline{x})}}_{\geqslant 0} d\underline{X}}_{\gtreqless 0} \qquad \cancel{\geq 0}$$

So D_{KL} can be +ve, −ve, 0

$$= \mathop{E}_{\underline{X} \sim p} \left[\ln \underbrace{\frac{\cancel{\ln}\ \cancel{g}\, p(\underline{X})}{g(\underline{X})}}_{g(\underline{X})} \right]$$

expectation over $p(\underline{x})$