Extra class

28 December 2020 20:18

Kernel machines:
Mercer's property
Given a kernel, if it satisfies m - there exists phi s.t phi^T phi = ker

- SVM: not necessarily linearly separable
- No theorem which says there always exists a kernel that can make the data linearly separable
- How do we choose the kernel?
- SVMs already regularized - hinge loss
  - Don't need to compute phi(x) w hingeloss bec dot products uwu
  - Wts - sum(yi,phi,lambda i) over support vecs
  - Kernel which to choose? We dunno - do a grid-search and choose the 1 with best validation score

Sigmoid: x-> inf, fn -> 1, and x_>-inf, fn->0 general defn, in the univ. approx thm paper

* Gradients & Subgradients
↳ for non diff able.

$h_{DNN} = W_3^T \phi(W_2^T \phi(W_1^T x)))$ — depth 3.

UAT doesn't tell no. of neurons

stacking w. height better than depth.

as decision bdry more complex no. neur exponentially reqr.

$\hat{R} = \sum_{i=1}^{n} y_i \log [h_{NN}(x_i)] \rightarrow$ [CE loss] $+ \alpha \Omega(W)$

Composite fn. with UAT. (Given sufficient.)

· Big advantage

Multilayer perceptron

CNNs :-

100 neurons but $x \in \mathbb{R}^{240 \times 450 \times 3}$

- Model complexity
- Imgs: Gridlike

Yann Le Cun 1994 → LeNet — Vala guy

↳ Linear don't care for topology

large neurons: overfitting ez ↳ No guarantee for generalizatn

Regularize
↓
Any regularization is ≡ some prior imposition on the parameter

CNN ≡ imposition of v. strong prior on MLP (Multilayer ...) → Grid like Topo...

Neuro Science ...

① ③ ML Estimate of P_x.

1 & 2

Why model unlabelled data? → To identify patterns in data

* Fitting a Mixture Model:

valid density fn. ← $P_\theta = \sum \alpha_j P_j(\theta_j)$ → Convex Combinatn

↳ a fn of x not θ, parametrized by θ_j

$\alpha_j \in [0,1]$, $\sum \alpha_j = 1$

$$P_\theta = \sum_{j=1}^{n} \alpha_j N(\mu_j, \Sigma_j)$$

MLE ≡ Min KL div. | $\theta = \{\alpha, \mu, \Sigma\}$

$$l(X, \alpha, \mu, \Sigma) = \sum_{i=1}^{n} \log(P_\theta x_i)$$

* EM: iterative algo → Maxm likelihood fn

$$l = \sum_n \log \sum_m \alpha_i N_j$$

→ if we knew which component x_i is coming from

↑ This is modelled as another RV

Latent RV

$x_i, z_i \sim P_z$ → prob. of sampling from a comp.

Mostly discrete (discrete for mixture densities)

Densities with nonoverlapping supports → Prob. a sample is from a particular comp 1 or 0.

2 stage sampling
2 lvls of randomness

$P_x = \int P_z \cdot P_{x|z} \, dz$] if

→ Model $P(z|x)$

· if know this then can max l.

that distrbn

N have ∞ support & any pt has nonzero prob. of belonging to every comp.

→ More belongingness closer to mean.

whats the prob. of x belonging to a particular comp?

k Mean, 1) $P(z|x)$ / 2) Max. L.

Neuro Science !

prior on MLP (Multilayer Perceptron) → Gridlike Topo. diff imposition

↓

Hubel & Wisel → specialization of neurons

force to spl. in up

[Compute $W^T x$

1) Local Receptive Field → [Map neuron to area of img.]

amt. of dimensions that every neuron gets to see.

m

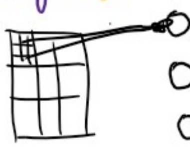
1 neuron

.neuron said to fire when > 0.5

Eg. detects edges

want it to be looking @ entire img

2) Parameter sharing →



$W_1^T x_{1:3}$

$W_1^T x_{10:19}$ clone the neuron ← {irrespective of where feature is

$W_1^T x_{n-9:n}$

↑

wts shared across all neurons.

[Kernel / filter / wt.] → W

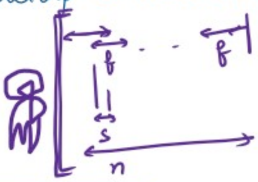
1 layer only

certain feat. detected.

3x3 filter.

apply on 3x3 parts of n.

Striding → overlap diff measure.



(f-s)

f f s n

↳ n is made mult. of f

pad. w.0

3) Subsampling / Pooling

* Grandmother neurons.

PxQ → KxL [avg of all pixels. / max Pooling

via subsampling

.LeNet →

32x32 → 5x5, 6 filter, Stride = 1 → 28x28x6

↓

5x5, 16

↓

10x10, 16

↓

Flatten

↓

MLP

1	0	-1
2	0	-2
1	0	-1

→ Sobel filter

60k params

AlexNet → New archi.

VGG Network

2 3x3 filter in series ≡ 1 5x5 filter

↓

K Means 1) P(Z|X) / 2) Max. L.

∴ Randomly assign x_i to one of K clusters. (1) $\equiv \alpha_i$s

. Compute → dist. b/w every x_i & each cluster centers.

. Reassign x_i to new clusters based on all dist. re estimate P(Z|X)

When $\sigma \to 0$

EM ≡ GMMs

Generative Models →

[P_Z ← sample Z then
P(X|Z) sample]

(1,0) → 1.23456

2 3x3 filter in series ≡ 1 5x5 filter

3x3 opt, replacing w. 5x5

keep

ResNet →

CNN →

Sobel filter

predef. filter

Convolve it

onv

RNN / LSTM → architectural prior for timeseries