Statistical Machine Learning

Christian Walder

College of Engineering and Computer Science
The Australian National University

Canberra

Semester One, 2020.

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Part X

Number of layers

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- expression of a function is compact when it has few computational elements, i.e. few degrees of freedom that need to be tuned by learning
- for a fi
  co
  bett
- Exa
  - affine operations, sigmoid $\Rightarrow$ logistic re
    1, fixed number of units (a.k.a. neurons)
  - fixed kernel, affine operations $\Rightarrow$ kern
    has two levels, with as many units as data points
  - stacked neural network of multiple "linear transformation followed by a non-linearity" $\Rightarrow$ deep neural network has arbitrary depth with arbitrary number of units per layer

Easier to represent with more layers

An old result:

- functions that can be compactly represented by a depth k architecture might require an exponential number of computational elements to be represented by a depth $k-1$ architecture
- Exa

pa 1 d 0 o

- Theorem: d-bit parity circuits of depth 2 have e size

Analogous in modern deep learning:

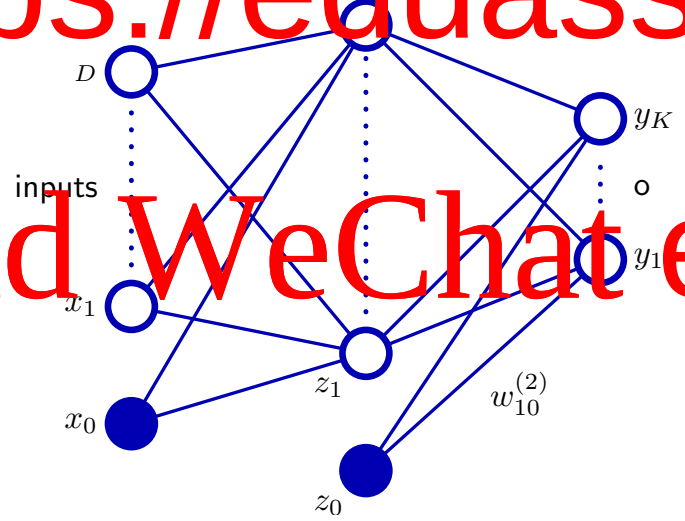
- "Shallow networks require exponentially more parameters for the same number of modes" — Canadian deep learning mafia.

Recall: Multi-layer Neural Network Architecture

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

$$y_k(\mathbf{x}, \mathbf{w}) = g\left(\sum_{j=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)$$

where $\mathbf{w}$



We could add more hidden layers

Empirical observations - pre 2006

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Deep architectures get stuck in local minima or plateaus
- As a gen
- Har
- 1 or 2 hidden layers seem to perform better
- 2006: Unsupervised pre-training, find distrib representation

Data61 | CSIRO
The Australian National University

Deep representation - intuition

Bengio, "Learning Deep Architectures for AI", 2009

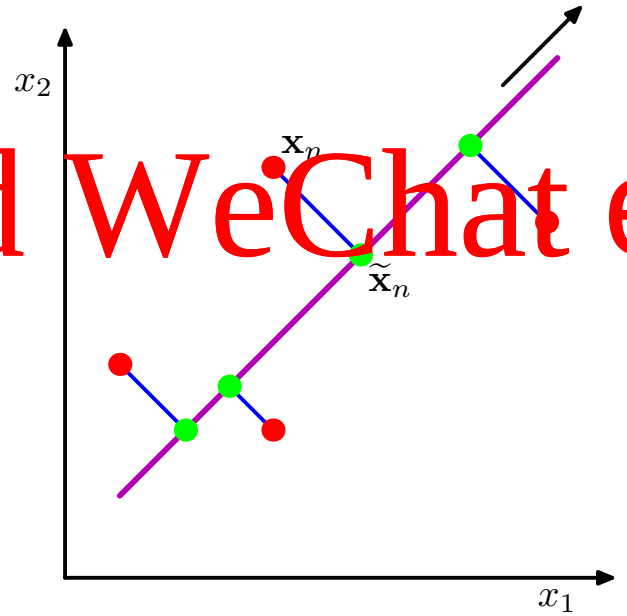
Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

AlexNet / VGG-F network visualized by mNeuron.

Recall: PCA

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Idea: Linearly project the data points onto a lower dimensional subspace such that
  - the variance of the projected data is maximised, or
  - the distortion error from the projection is minimised
- Both formulation lead to the same result.
- Ne
  prin



Multiple PCA layers? - linear transforms

Data61 | CSIRO
The Australian National University

- Principle Component Analysis is a linear transformation (because it is a projection)
  - The
  - Lin
  - Let that

$$ST(X + X') = ST(X) + S$$

  - Similarly for multiplication with a scalar

$\Rightarrow$ multiple PCA layers pointless

Multiple PCA layers? - projection

- Let $X^T X = U \Lambda U^T$ be the eigenvalue decomposition of the covariance matrix (what is assumed about the mean?).
- Define U_k to be the matrix of the first k columns of U, for the k largest eigenvalues. Define Λ_k similarly
- Co
prin

- We perform PCA a second time, $Z^T Z$
- By the definition of Z and $X^T X$ and the ort

$$Z^T Z = (XU_k)^T (XU_k) = U_k^T X^T X U_k = U_k \, U \Lambda U \;\; U_k = \Lambda_k$$

- Hence $\Lambda_Z = \Lambda_k$ and V is the identity, therefore the second PCA has no effect

$\Rightarrow$ again, multiple PCA layers pointless

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Autoencoder

- An autoencoder is trained to encode the input x into some representation $c(x)$ so that the input can be reconstructed from that representation
- the target output of the autoencoder is the autoencoder input itself
- Wit
  crit
  spa
- If the hidden layer is nonlinear, the autoencode differently from PCA, with the ability to capture m aspects of the input distribution
- Let f be the decoder. We want to minimise the reconstruction error

$$\sum_{n=1}^{N} \ell\left(x_n, f(c(x_n))\right)$$

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Cost function

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Recall: $f(c(x))$ is the reconstruction produced by the network
- Minimisation of the negative log likelihood of the reconstruction, given the encoding $c(x)$
- If $x|$
- If the inputs x_i are either binary or considere binomial probabilities, then the loss function is cross entropy

$$-\log P(x|c(x)) = -x_i \log f_i(c(x)) + (1 - x_i) \log(1 - f_i(c(x)))$$

where $f_i(\cdot)$ is the i^{th} component of the decoder

Undercomplete Autoencoder

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Co
- $c(x)$
- Ca
  exa
- Hope code $c(x)$ is a distributed representa
  captures the main factors of variation in the data

Stacking autoencoders

- Let c_j and f_j be the encoder and corresponding decoder of the j^{th} layer
- Let z_j be the representation after the encoder c_j
- We can define multiple layers of autoencoders recursively.
- For the c
- Bec feature z_2 can capture more complex patter

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Pre-training supervised neural networks

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Latent features z_j in layer j can capture high level patterns

$$z_j = c_j(c_{j-1}(\quad c_2(c_1(x))\quad))$$

- The tas
- In co constructed in an unsupervised fashion. j
- Discard the decoding layers, and directly use supervised training method, such as logistic re
- Various such pre-trained networks are available on-line, e.g VGG-19.

Xavier Initialisation / ReLU

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Layer-wise unsupervised pre-training helps by extracting useful features for subsequent supervised backprop.
- Pre-training also avoids saturation (large magnitude arguments to, say, sigmoidal units).
- Simpler Xavier initialization can also avoid saturation.
- Let t
acti

$$\text{VAR}[z] = \mathbb{E}[(z - \mathbb{E}[z])^2] = \mathbb{E}[z^2] = \qquad ^2$$

$$= \sum_{i=1}^{m} \mathbb{E}[(x_i w_i)^2] = \sum_{i=1}^{m} \mathbb{E}[x_i$$

- So we set $\sigma = 1/\sqrt{m}$ to have "nice" activations.
- Glorot initialization takes care to have nice back-propagated signals — see the auto-encoder lab.
- ReLU activations $h(x) = \max(x, 0)$ also help in practice.

Higher dimensional hidden layer

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- if there is no other constraint, then an autoencoder with d-di
  d c
- Av

  - Early stopping of stochastic gradient descen
  - Add noise in the encoding
  - Sparsity constraint on code $z(x)$.

Denoising autoencoder

- Add noise to input, keeping perfect example as output
- Aut
  1.
  2.
- Re

$$-\log P(x|c(\hat{x}))$$

where x noise free, $\hat{x}$ corrupted

Image denoising

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- Images with Gaussian noise added.

- Denoised using Stacked Sparse Denoising A

Images from Xie et. al. NIPS 2012

Inpainting - 1

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Image from `http://cimg.eu/greycstoration/demonstration.shtml`

Inpainting - 2

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

Undo text over image

Image from Bach et. al. ICCV tutorial 2009

Recall: Basis functions

©2020 Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- For fixed basis functions $\phi(x)$, we use domain knowledge for encoding features
- Ne
  $\phi_i(x$
  lear
- The transformations $\phi_i(\cdot)$ for a particular d
  longer be orthogonal, and furthermore may be
  variations of each other.
- We collect all the transformed features into a ma

Sparse representations

- Idea: Have many hidden nodes, but only a few active for a particular code $c(x)$
- Student t prior on codes
- ℓ_1 penalty on coefficients α

$$\min_{\alpha} \sum_{n=1}^{N} \frac{1}{2} \|x_n - \Phi \alpha_n\|_2^2$$

- Φ is overcomplete, no longer orthogonal
- Sparse $\Rightarrow$ small number of non-zero α_i.
- Exact recovery under certain conditions (coherence): $\ell_1 \to \ell_0$.
- ℓ_1 regulariser $\sim$ Laplace prior $p(\alpha_i) = \frac{\lambda}{2} \exp(-\lambda |\alpha_i|)$.

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

$$\underbrace{y}_{\text{measurements}} = \underbrace{x_{orig}}_{\text{original image}} + \underbrace{\quad}_{\text{noise}}$$

Sparsity assumption

Data61 | CSIRO
The Australian National University

- Only have noisy measurements

$$y = x_{ori} + w$$

measurements original image noise

- Giv

where $\|\cdot\|_0$ is the number of non-zero eleme

- Φ is not necessarily features constructed fro data.

- Minimise reconstruction error

$$\min_{\alpha} \sum_{n=1}^{N} \frac{1}{2} \|x_n - \Phi\alpha_n\|_2^2 + \lambda\|\alpha\|_0$$

Convex relaxation

- Want to minimise number of components

$$\min_{\alpha} \sum_{n=1}^{N} \frac{1}{2}\|x_n - \Phi\alpha_n\|_2^2 + \lambda\|\alpha\|_0$$

- but
- Rel

$$\min_{\alpha} \sum_{n=1}^{N} \frac{1}{2}\|x_n - \Phi\alpha_n\|_2^2 +$$

where $\|\alpha\|_1 = \sum_n |\alpha_n|$.

- In some settings does minimisation with ℓ_1 regularisation give the same solution as minimisation with ℓ_0 regularisation (exact recovery)?

Mutual coherence

- Expect to be ok when columns of Φ "not too parallel"
- Assume columns of Φ are normalised to unit norm
- Let
  of th
- De

$$M = M(\Phi) = \max_{i \neq j} |K(i$$

- If we have an orthogonal basis, then Φ matrix, hence $K(i,j) = 0$ when $i \neq j$.
- However, if we have very similar columns, then $M \approx 1$.

Exact recovery conditions

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

- If a minimiser of the true ℓ_0 problem α^* satisfies

$$\frac{1}{}$$

the

- If α

$$\|\alpha^*\|_0 < \frac{1}{2}\left(1 + \frac{1}{M}\right)$$

then the minimiser of the ℓ_1 relaxation h
sparsity pattern as α^*.

Motivation

Autoencoder

References

- Yoshua Bengio, "Learning Deep Architectures for AI", Foundations and Trends in Machine Learning, 2009
- ht
- ht au
- Fuchs, "On Sparse Representations in Arbitrary Redundant Bases", IEEE Trans. Info. Theory
- Xavier Glorot and Yoshua Bengio, Underst difficulty of training deep feedforward neural n 2010.