Intelligent Systems, Revision

Pierre Gaillard

April, 2023

INRIA

Hyperparameters

1. What is the difference between a parameter and an hyper-parameter?
2. Name the hyper-parameters of the following methods:

- KNN
- Logistic regression, with ℓ_2 regularization
- SVM with soft-margin
- Multi-layer perceptron
- Deep Convolutional Neural Network
- PCA
- Decision Tree

3. For each hyper-parameter, tell if you increase the parameter if it leads to more, same, or less regularization (more bias, less variance).
4. What is the standard solution to tune hyper-parameters?

Optimization

Often, to train algorithms, one needs to minimize an empirical loss

$$\widehat{\theta} \in \underset{w \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, f_\theta(x_i)\big)$$

1. Name a few algorithms
2. When does overfitting may occur?
3. How to avoid it on such algorithms?
4. How do you solve such a problem?

Optimization

Often, to train algorithms, one needs to minimize an empirical loss

$$\widehat{\theta} \in \underset{w \in \mathbb{R}^d}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, f_\theta(x_i)\big)$$

1. Name a few algorithms (Linear/Logistic/Ridge/Lasso regression, Neural Networks, ...)
2. When does overfitting may occur?
3. How to avoid it on such algorithms?
4. How do you solve such a problem?

PCA vs. t-SNE

What is the difference between PCA and t-SNE?

Draw and explain an example where t-SNE work better than PCA.

PCA vs. t-SNE

What is the difference between PCA and t-SNE?

Draw and explain an example where t-SNE work better than PCA.

Comparison between the methods

PCA has a 2 big advantages compared to t-SNE:

- It is **deterministic**
- the axis are **interpretable** as they are a linear combination of the variables (cf. stat lectures).
- no parameter to tune (target entropy in case of t-SNE)

t-SNE has the advantage at looking only at local scale, which is often relevant, and is **non-linear** projection method.

PCA vs. t-SNE

What is the difference between PCA and t-SNE?

Draw and explain an example where t-SNE work better than PCA.

Exemple where t-SNE is better than PCA

If the samples are primarily similar to close neighbors, and the large distances between samples are less important, t-SNE can work better than PCA. A typical example is the spiral, where long distance is “irrelevant” and neighbor connectivity is important.

Conditional Log-Likelihood

What is conditional log-likelihood?

1. Assume that $X|Y = i \sim \mathcal{N}(\mu_i, I)$.
   a) What is the parameter to be estimated?
   b) Wat is the conditional log-likelihood in this case?
   c) Show that it can be written as $\sigma(w \cdot \varphi(x) + b)$. Determine w and b
3. What happens if $X|Y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$?

Ridge regression

Let $(x_i, y_i)_{i=1,\dots,n}$ be some data in $\mathbb{R}^d \times \mathbb{R}$.

1) What problem does Ridge regression solves here?
2) Write the closed form expression of the parameter.
3) Assume that the data looks like (draw on blackboard), would it be a good algorithm? What could you use to fix it?

Activation function

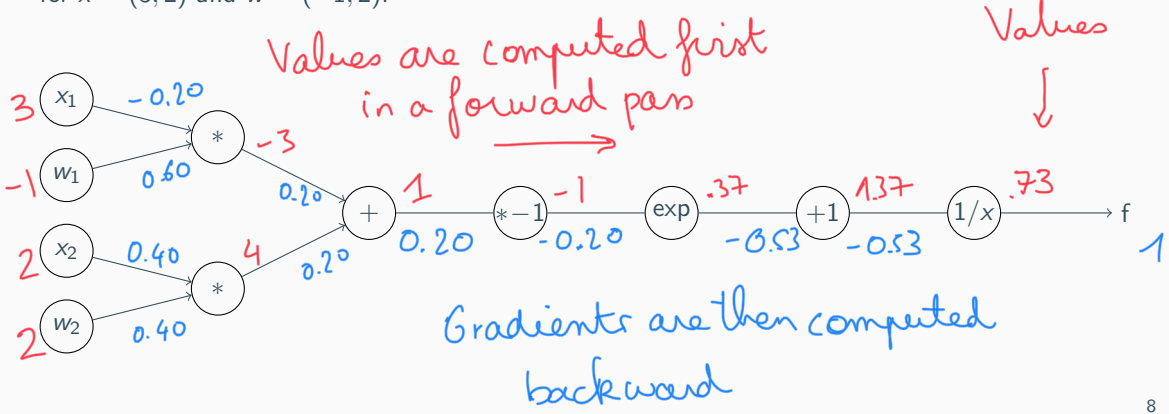
What is the sigmoid activation function? What are the problems with using a sigmoid activation function?

Back-propagation

Perform back-propagation on the following simple Neural Network that computes

$$f(x) = \frac{1}{1 + \exp(-(w_1 x_1 + w_2 x_2))}$$

for $x = (3, 2)$ and $w = (-1, 2)$.

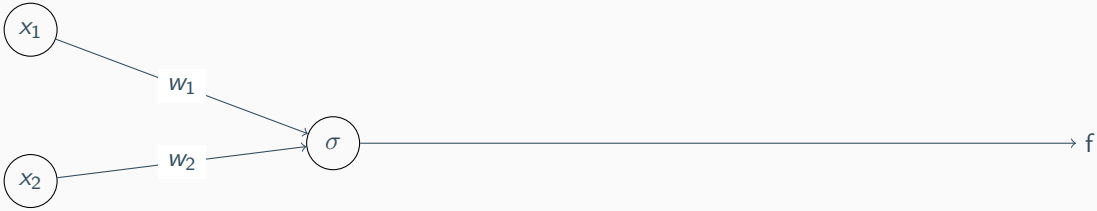


Back-propagation

Perform back-propagation on the following simple Neural Network that computes

$$f(x) = \frac{1}{1 + \exp(-(w_1 x_1 + w_2 x_2))} = \sigma(w_1 x_1 + w_2 x_2)$$

for $x = (3, 2)$ and $w = (-1, 2)$.



CNN

You have a 32x32x5 image and filter it with a 5x5x5 kernel, the way most convolutional neural networks are implemented. If you use no padding, what will be the output size of the activation map?

CNN

You have a 32x32x5 image and filter it with a 5x5x5 kernel, the way most convolutional neural networks are implemented. If you use no padding, what will be the output size of the activation map?

28x28x1

Scores

What is the difference between Recall, Precision, and Accuracy?
Compute them for the following confusion matrix (on black-board).

Gaussian Mixture Models vs. Kmeans

What is the difference between Gaussian Mixture Model and Kmeans?

Draw examples of data.

How do you solve Gaussian Mixture Model?

References

Thank you!