5 - Deep Learning Overview

5.4 - Regularization

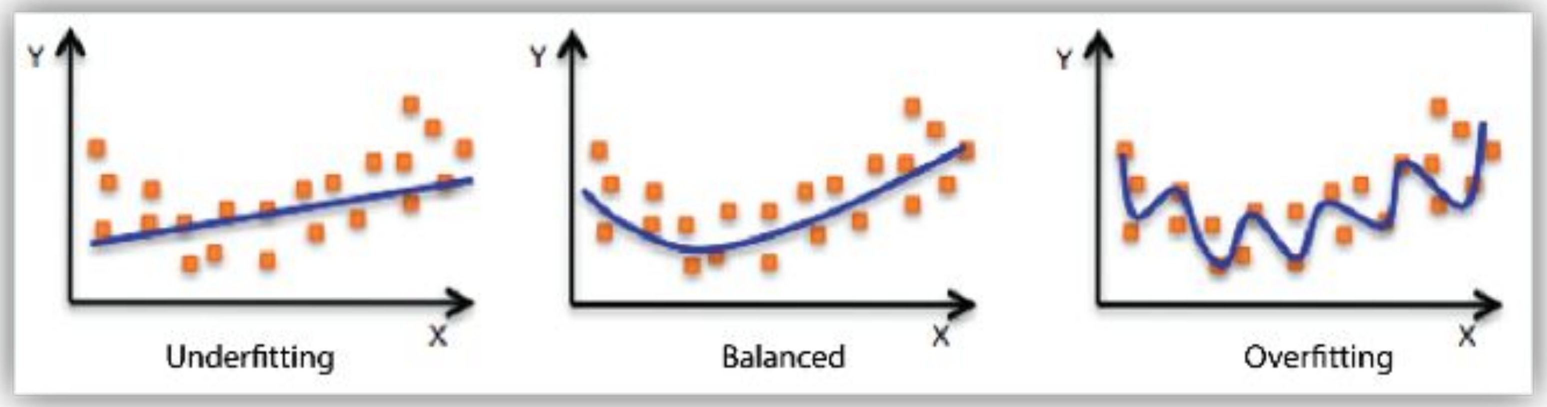
Marciel Barros

Setembro, 2020

1. Deep Learning Summary
2. Training Rules
3. Activation Function
4. **Regularization**
5. Optimizer
6. Types of Neural Networks
7. Common Problems

Overfitting

- Overfitting occurs when network parameters are not good to generalize input data;



Regularization

- Regularization is an important and effective technology to reduce generalization errors in machine learning. It is especially useful for deep learning models that tend to be overfit due to a large number of parameters. Therefore, researchers have proposed many effective technologies to prevent overfitting, including:
  - Adding constraints to parameters, such as L_1 and L_2 norms
  - Expanding the training set, such as adding noise and transforming data
  - Dropout
  - Early-stopping

Penalty Parameters

- Many regularization methods restrict the learning capability of models by adding a penalty parameter $\Omega(\theta)$ to the objective function J. Assume that the target function after regularization is $\tilde{J}$.

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta),$$

Where $\alpha \epsilon [0, \infty)$ is a hyperparameter that weights the relative contribution of the norm penalty term Ω and the standard objective function $J(X; \theta)$. If α is set to 0, no regularization is performed. The penalty in regularization increases with α.

L_1 Regularization

- Add L_1 norm constraint to model parameters, that is,

$$\tilde{J}(w; X, y) = J(w; X, y) + \alpha \|w\|_1,$$

- If a gradient method is used to resolve the value, the parameter gradient is

$$\nabla \tilde{J}(w) = \propto sign(w) + \nabla J(w).$$

L_2 Regularization

- Add norm penalty term L_2 to prevent overfitting.

$$\tilde{J}(w; X, y) = J(w; X, y) + \frac{1}{2}\alpha\|w\|_2^2,$$

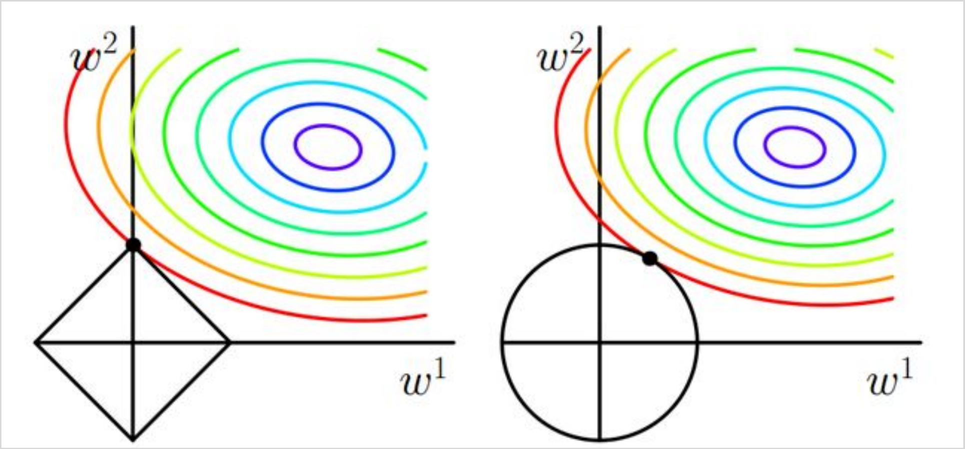
A parameter optimization method can be inferred using an optimization technology (such as a gradient method):

$$w = (1 - \varepsilon\alpha)\omega - \varepsilon\nabla J(w),$$

where ε is the learning rate. Compared with a common gradient optimization formula, this formula multiplies the parameter by a reduction factor.

L_2 vs L_1

- The major differences between L_2 and L_1:
  - According to the preceding analysis, L_1 can generate a more sparse model than L_2. When the value of parameter w is small, L_1 regularization can directly reduce the parameter value to 0, which can be used for feature selection.
  - From the perspective of probability, many norm constraints are equivalent to adding prior probability distribution to parameters. In L_2 regularization, the parameter value complies with the Gaussian distribution rule. In L_1 regularization, the parameter value complies with the Laplace distribution rule.

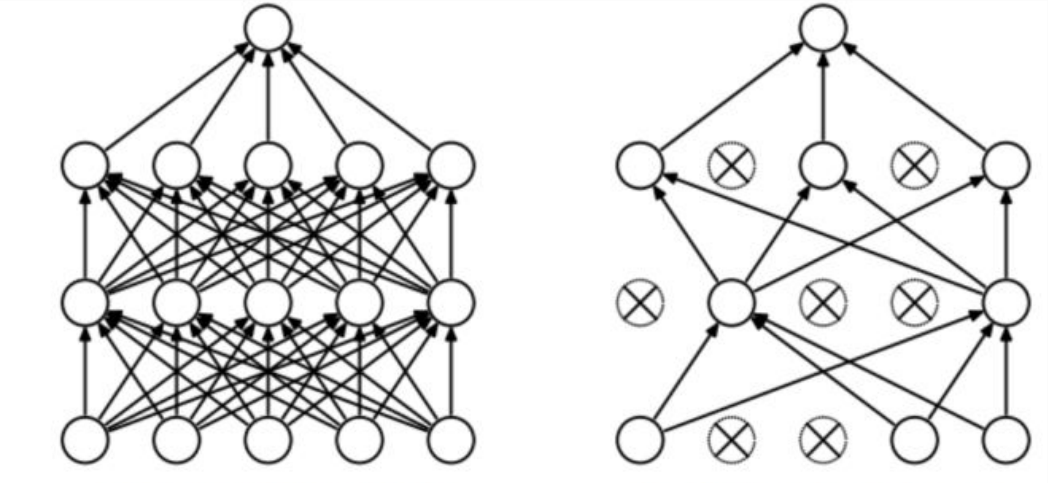


Dataset Expansion

- The most effective way to prevent overfitting is to add a training set. A larger training set has a smaller overfitting probability. Dataset expansion is a time-saving method, but it varies in different fields.
  - A common method in the object recognition field is to rotate or scale images. (The prerequisite to image transformation is that the type of the image cannot be changed through transformation. For example, for handwriting digit recognition, categories 6 and 9 can be easily changed after rotation).
  - Random noise is added to the input data in speech recognition.
  - A common practice of natural language processing (NLP) is replacing words with their synonyms.
  - Noise injection can add noise to the input or to the hidden layer or output layer. For example, for Softmax classification, noise can be added using the label smoothing technology. If noise is added to categories 0 and 1, the corresponding probabilities are changed to $\frac{\varepsilon}{k}$ and $1-\frac{k-1}{k}\varepsilon$ respectively.

Dropout

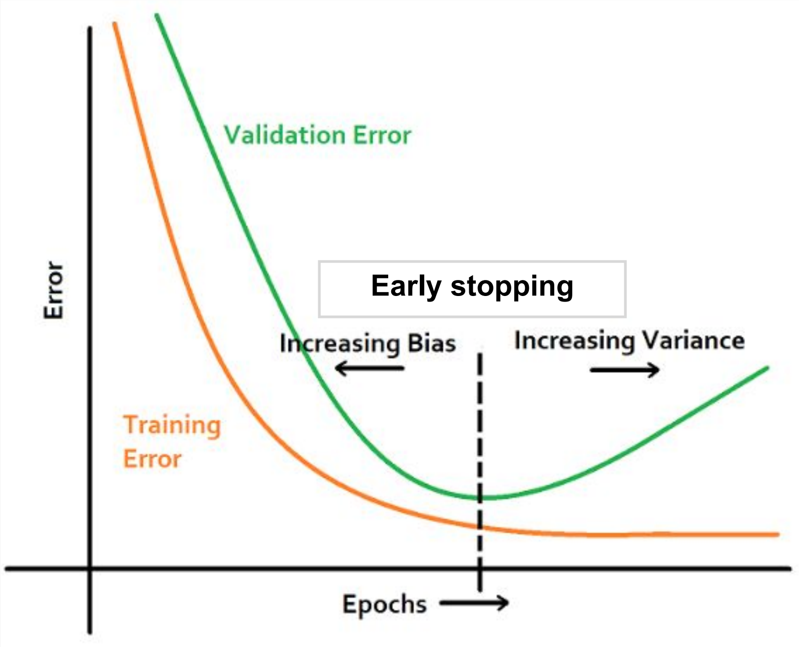
- Dropout randomly discards some inputs during the training process. In this case, the parameters corresponding to the discarded inputs are not updated.
- As an integration method, Dropout combines all sub-network results and obtains sub-networks by randomly dropping inputs.



(a) Standard Neural Net (b) After applying dropout.

Early Stopping

- A test on data of the validation set can be inserted during the training. When the data loss of the verification set increases, perform early stopping.



5 - Deep Learning Overview

Next: 5.5 - Optimization for Model Training

Marciel Barros

Setembro, 2020