1 Chapter 4 Numerical Optimization

Problems of under- and over-flow.

Find x that minimizes the criterion function $f(x)$.

- Gradient descent method as an example of first-order optimization.
- Using Hessian matrix, we have second order optimization, that's related to the eigenvalue-eigenvector of the Hessian matrix.
- KKT condition.

2 Chapter 5 Basics of Machine Learning

5.7, 5.8 on supervised and unsupervised learning.

3 Chapter 6 DNN

Deep feedforward networks has no feedback. We use the composition of a network of functions $f(x)$ to approxiamte $f^*(x)$, our training set is **training points** and **label** $y \sim f^*(x)$. The final layer is **output layer** other layers are **hidden layers**.

We can generalize the linear models by applying liear model to transformed input $\phi(x)$, where ϕ is a set of features describing x or a new representation of x.

Deep learning tries to learn ϕ: $y = \phi(x;\theta)'w$.

Example 3.1. Learn the XOR function.

3.1 6.2 How to choose cost function and output layer

Given training set (x, y) where x is a vector, loss function from likelihood analysis is

$$J(\theta) = -E_{\text{sample}} \log P_{\text{model}}(y \mid x)$$

What we learn from the training set is not $\hat{y}$ that approximates y, but the conditional distribution. Regression is about the mean of the conditional distribution. If we use L^2 loss, then we can learn the mean, if we use L^1 loss we can learn the median.

The output layer design usually uses the following functions:

1. suppose y is normal, then a *linear* output layer is used, commonly for regression problem.

2. For Bernoulli distribution, we can use sigmoid:

$$\frac{1}{1+e^{-x}}$$

   when we use log likelihood, we can avoid the problem of having a derivative that's too close to 0 and achieve better numerical performance.. The overparametrization problem won't matter in practice.

3. For classification, most commonly used output layer is softmax. The *overparametrization* problem won't matter in practice.

Sometimes we consider Gaussian Mixture model, but then we need to estimate the covariance matrices and the results aree usually unstable because we need matrix division?

3.2 6.3 How to Choose Hidden Layers

So far there isn't a theory guidance on that. Commonly used hidden layer include the ReLU. Other more complicated choices don't seem to perform better in empirical studies.