Optimization Objectives

Amaires@May 2024

1 Introduction

Given samples (x, y) from a distribution with probability density function $p(x, y)$, the optimization goal of a classification problem or a regression problem is to find a good $p_\theta(y|x)$ where θ is the parameter of a chosen family of probability density functions. The objective can be derived in three different but related ways.

1.1 K-L divergence of conditional distribution

One criterion of a good $p_\theta(y|x)$ is how close it is to $p(y|x)$. One such closeness measure is the K-L divergence between $p(y|x)$ and $p_\theta(y|x)$ which is $\int p(y|x) \log \frac{p(y|x)}{p_\theta(y|x)} dy$. Of course, this should work across all x, therefore our objective should be

$$\min_\theta \int p(x)[\int p(y|x) \log \frac{p(y|x)}{p_\theta(y|x)} dy]dx = \min_\theta [\int p(x)[\int p(y|x) \log p(y|x)dy]dx - \int p(x)[\int p(y|x) \log p_\theta(y|x)dy]dx]$$

For our purpose, the first term is an unknown consant independent of θ. Removing this constant, our objective changes to

$$\begin{aligned}\min_\theta - \int p(x)[\int p(y|x) \log p_\theta(y|x)dy]dx &= \min_\theta -E_{x\sim p(x)}E_{y\sim p(y|x)} \log p_\theta(y|x) \\ &= \min_\theta -E_{x,y\sim p(x,y)} \log p_\theta(y|x)\end{aligned}$$

The left side of the above equation can also be written as:

$$\begin{aligned}\min_\theta - \int p(x)[\int p(y|x) \log p_\theta(y|x)dy]dx &= \min_\theta - \int \int p(x)p(y|x) \log p_\theta(y|x)dydx \\ &= \min_\theta - \int \int p(x, y) \log p_\theta(y|x)dydx \\ &= \min_\theta - \int p(y) \int p(x|y) \log p_\theta(y|x)dxdy \\ &= \min_\theta -E_{y\sim p(y)}E_{x\sim p(x|y)} \log p_\theta(y|x) \\ &= \min_\theta -E_{x,y\sim p(x,y)} \log p_\theta(y|x)\end{aligned}$$

So basically, the optimization objective is the following three equivalent functions:

$$\begin{aligned}&\min_\theta -E_{x\sim p(x)}E_{y\sim p(y|x)} \log p_\theta(y|x) \\ &\min_\theta -E_{y\sim p(y)}E_{x\sim p(x|y)} \log p_\theta(y|x) \\ &\min_\theta -E_{x,y\sim p(x,y)} \log p_\theta(y|x)\end{aligned}$$

1.2 K-L divergence of joint distribution

Since $p_\theta(x, y) = p(x)p_\theta(x, y)$, it is easy to arrive at the same conclusions by minimizing the K-L divergence between $p(x, y)$ and $p_\theta(x, y)$:

$$\begin{aligned}\min_\theta \int \int p(x, y) \log \frac{p(x, y)}{p_\theta(x, y)} dxdy &= \min_\theta \int \int p(x, y) \log \frac{p(x)p(y|x)}{p(x)p_\theta(y|x)} dxdy \\ &= \min_\theta \int \int p(x, y) \log \frac{p(y|x)}{p_\theta(y|x)} dxdy \\ &= \min_\theta [\int \int p(x, y) \log p(y|x)dxdy - \int \int p(x, y) \log p_\theta(y|x)dxdy]\end{aligned}$$

Again, the first term is an unknown constant independent of θ that can be removed. The objective changes to

$$\begin{aligned}\min_\theta - \int \int p(x, y) \log p_\theta(y|x)dxdy &= \min_\theta E_{x,y\sim p(x,y)} \log p_\theta(y|x) \\ &= \min_\theta \int p(x)[\int p(y|x) \log p_\theta(y|x)dy]dx = \min_\theta E_{x\sim p(x)}E_{y\sim p(y|x)} \log p_\theta(y|x) \\ &= \min_\theta \int p(y)[\int p(x|y) \log p_\theta(y|x)dx]dy = \min_\theta E_{y\sim p(y)}E_{x\sim p(x|y)} \log p_\theta(y|x)\end{aligned}$$

1.3 Maximum likelihood

Given a set of samples (x_i, y_i), assumed to be i.i.d, one objective could be to maximize the likelihood of observing these samples, which is

$$\max_\theta \prod_i p_\theta(x_i, y_i)$$

This is equivalent to minimizing the negative log likelihood

$$\begin{aligned}
\min -\sum_i \log p_\theta(x_i, y_i) &= \min_\theta -\sum_i \log p(x_i)p(y_i|x_i) \\
&= \min_\theta -(\sum_i \log p(x_i) + \sum_i \log p_\theta(y_i|x_i))
\end{aligned}$$

As before, the first term is an unknown constant independent of θ. Once the first term is removed, the ojective becomes

$$\min_\theta -\sum_i \log p_\theta(y_i|x_i)$$

Divide it by the number of samples, and rewirte it in expectation form, the objective becomes

$$\min_\theta -E_{x,y\sim p(x,y)} \log p_\theta(y|x)$$

This is the same as what is derived in Section 1.1 and Section 1.2.

2 Classification

In a classification problem, y takes on a fixed number of possible values usually encoded using numbers from 1 through K. A classifier usually outputs the entire probability vector $p_\theta(y = 1|x), p_\theta(y = 2|x), p_\theta(y = 3|x), \cdots, p_\theta(y = K|x)$. In the case of a binary classification problem, however, it is more customary to use $\{0, 1\}$ to encode the two possible values that y can take, and the classifier only outputs $f(x) = p_\theta(y = 1|x)$ with $p_\theta(y = 0|x)$ implied to be $1 - f(x)$. In this case, the optimization objective can be rewritten as

$$\min_\theta -E_{x,y\sim p(x,y)}(y \log f(x) + (1 - y) \log(1 - f(x))$$

This is usual called the binary cross entropy objective.

3 Regression

In a regression problem, a neural network's output can be interpreted as the mean $\mu_\theta(x)$ of a normal distribution $N(\mu_\theta(x), I)$. With this interpretation, the optimization objective can be rewritten as

$$\begin{aligned}
\min_\theta -E_{x,y\sim p(x,y)} \log p_\theta(y|x) &= \min_\theta -E_{x,y\sim p(x,y)} \log(2\pi)^{-\frac{d}{2}} \exp(-\frac{1}{2} \|y - \mu_\theta(x)\|^2) \\
&= -\frac{d}{2}\log(2\pi) + \frac{1}{2} \min_\theta E_{x,y\sim p(x,y)} \|y - \mu_\theta(x)\|^2
\end{aligned}$$

where d is the dimension of y. This objective is equivalent to

$$\min_\theta E_{x,y\sim p(x,y)} \|y - \mu_\theta(x)\|^2$$

which is the well known mean squared error objective.