Homework 8 Solution

Deep Learning

1 Regression

The posterior probability of Θ is $p(\Theta|D)$, and from the Bayes rule we have

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)} \propto p(D|\Theta)p(\Theta) \tag{1}$$

where we have used the fact that $p(D)$ is constant for any Θ. Maximizing the posterior probability of Θ is equivalent to maximizing the logarithm of the posterior, i.e.,

$$\begin{aligned}\Theta^* &= \underset{\Theta}{\operatorname{argmax}}\ \log p(\Theta|D) &(2)\\ &= \underset{\Theta}{\operatorname{argmax}}\ \log p(D|\Theta) + \log p(\Theta)\ . &(3)\end{aligned}$$

Now for $p(D|\Theta)$ with the i.i.d. assumption for the data we have

$$\begin{aligned}p(D|\Theta) &= \prod_{i=1}^{M} p(X_i, Y_i|\Theta) &(4)\\ &= \prod_{i=1}^{M} p(Y_i|X_i,\Theta)p(X_i|\Theta) &(5)\\ &= \prod_{i=1}^{M} p(Y_i|X_i,\Theta)p(X_i)\ , &(6)\end{aligned}$$

where the last equality holds since input X_i is independent of the parameters. By substituting $p(D|\Theta)$ from (6) into (3) and accounting for the independence of Θ from $p(X_i)$ we obtain

$$\begin{aligned}\Theta^* &= \underset{\Theta}{\operatorname{argmax}}\ \log p(D|\Theta) + \log p(\Theta) &(7)\\ &= \underset{\Theta}{\operatorname{argmax}}\ \log p(\Theta) + \sum_{i=1}^{M} \log p(Y_i|X_i,\Theta)\ . &(8)\end{aligned}$$

Now by replacing $p(\cdot)$ with the given Gaussian distributions, we have

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}}\ \left(\frac{\|\Theta\|^2}{2} + \sum_{i=1}^{M}\left[\log\sigma^2(X_i,\Theta) + \frac{(Y_i - f(X_i,\Theta))^2}{2\sigma^2(X_i,\Theta)}\right]\right)\ . \tag{9}$$

For estimating the variance of $Y|X$, we note that $Y|X \sim N(f(X,\Theta), \sigma^2(X,\Theta))$. We can follow two approaches to estimate the variance. In one approach, we can use a single neural network with two outputs $f(X,\Theta)$ and $\sigma^2(X,\Theta)$ and solve the optimization problem in (9) to find the solution for Θ, and the variance will be $\sigma^2(X,\Theta^*)$. In the second approach, we can use two separate neural networks, one for the mean $f(X,\Theta)$ and one for the variance $\sigma^2(X,\beta)$. Note that the parameters of these two neural networks are different. We can assume that the prior distribution for β is also Gaussian, i.e., $\beta \sim N(0,I)$. Then, we can solve for Θ and β by alternating between two optimization problem as described in [1]. If you describe any of these approaches, you get the full grade for this part.

[1] G. Papadopoulos, P. J. Edwards and A. F. Murray, "Confidence estimation methods for neural networks: A practical comparison," *IEEE Transactions on Neural Networks,* vol. 12, no. 6, pp. 1278-1287, Nov 2001.

2 Classification

For classification, if we estimate Θ by maximizing the posterior probability, we again have

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \; \log p(\Theta) + \sum_{i=1}^{M} \log p(Y_i|X_i,\Theta) \; . \tag{10}$$

By assuming the Bernoulli distribution for $Y|X$ we have

$$p(Y_i|X_i,\Theta) = [\sigma(f(X_i,\Theta))]^{Y_i}[1-\sigma(f(X_i,\Theta))]^{1-Y_i} \; . \tag{11}$$

Therefore,

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \; \left(\frac{\|\Theta\|^2}{2} + \sum_{i=1}^{M} [Y_i \log \sigma(f(X_i,\Theta)) + (1-Y_i)\log(1-\sigma(f(X_i,\Theta))] \right) \; . \tag{12}$$

Since the variance of a Bernoulli random variable with parameter α is $\alpha(1-\alpha)$, the variance can be estimated as

$$Var(Y|X) = \sigma(f(X,\Theta^*))(1-\sigma(f(X,\Theta^*)) \; . \tag{13}$$

Each problem has **25 points**; **15 points** for finding the objective function and **10 points** for the variance. For the objective function, you will receive **5 points** for using the Bayes rule correctly, **5 points** for identifying where you can use the independence of the random variables, and **5 points** for simplifying the results.