Homework Assignment 3
Radial Basis Function Networks and Backpropagation

Kailin Zheng kz882

March 27, 2019

1 RBFN for XOR problem (20 points)

According to course notes Page 21, with the four basis vectors r^1, r^2, r^3, r^4, we will transform each two-dimensional input vector x into a four-dimensional vector

$$\phi_i(x) = exp(-(x - r^i)^2).$$

The function's output is bounded between 0 and 1. The output is closer to 1 when the input vector x is close to the basis vector r, but converges to 0 as the distance between them grows.

That is to say, if the input vector x is closer to r^1, r^2, we tend to classify them as positive, so we set the weight to be 1 correspondingly; if the input vector x is closer to r^3, r^4, we tend to classify them as negative, so we set the weight to be -1 correspondingly.

If we construct the weight vector as $w = [1, 1, -1, -1, 0]^T$, where the first four elements represent the label of each basis vector, we got perceptron in a layered neural network. The machine is written as

$$M(x) = \sigma(w^T \phi(x) + b).$$

(The reason that we have a 0 as the fifth element in the weight vector is that we would like a bias term.)

$M(x) = \sigma(exp(-(x-r^1)^2) + exp(-(x-r^2)^2) - exp(-(x-r^3)^2) - exp(-(x-r^4)^2))$

If x is in 1 or 3 dimension, closer to r^1, r^2, the thing inside sigmoid function $>= 0$.

If x is in 2 or 4 dimension, closer to r^3, r^4, the thing inside sigmoid function < 0.

$$M(x) = \begin{cases} 1, & \text{if } w^T\phi(x) + b >= 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

2 Adaptive RBFN (20 points)

In addition to the distance function of logistic regression, we have feature engineering to the input x so:

$$\Delta(y^*, M, \phi(x)) = -(y^* \log M(\phi(x)) + (1 - y^*) \log(1 - M(\phi(x))))$$

First, we know $M(x) = \sigma(w^T\phi(x) + b)$, use it in:

$$\frac{\partial \Delta}{\partial a} = -(y^* - M(\phi(x))) = -(y^* - \sigma(w^T\phi(x) + b))$$

Second, we know $a = w^T\phi(x) + b$, use it in:

$$\frac{\partial a}{\partial \phi_k(x)} = w^T$$

Third, we know $\phi_k(x) = exp(-(x - r^k)^2)$, $\frac{de^{-x}}{dx} = -e^{-x}$ use it in:

$$\nabla_{r^k}\phi_k(x) = \nabla_{r^k} exp(-(x - r^k)^2) = -exp(-(x - r^k)^2)\nabla_{r^k}(x - r^k)^2$$
$$= -\phi_k(x) * (2(x - r^k)) * (-1) = 2\phi_k(x)(x - r^k)$$

Then we multiply the three gradients according to chain rule to get the ultimate gradient:

$$\begin{aligned} \nabla_{r^k}\Delta(y^*, M, \phi(x)) =& \underbrace{-(y^* - \sigma(w^\top\phi(x) + b))}_{\frac{\partial \Delta}{\partial a}} \underbrace{w_k}_{\frac{\partial a}{\partial \phi_k(x)}} \underbrace{(2\phi_k(x))(x - r^k)}_{\nabla_{r^k}\phi_k(x)} \\ =& -2(y^* - \sigma(w^\top\phi(x) + b))w_k\phi_k(x)(x - r^k) \end{aligned}$$

3 Probability quiz (20 points)

1. Show Bayes' rule is true:

Assume two events X and Y:

$$\begin{aligned} P(Y \cap X) &= P(Y)P(X|Y) \\ P(Y \cap X) &= P(X)P(Y|X) \\ P(Y)P(X|Y) &= P(X)P(Y|X) \\ P(Y|X) &= \frac{P(X|Y)P(Y)}{P(X)} \end{aligned}$$

2. Show linearity of expectation for discrete random variables X and Y:

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_x \sum_y (x+y)P(X=x, Y=y) \\
&= \sum_x x \sum_y P(X=x, Y=y) + \sum_y y \sum_x P(X=x, Y=y) \\
&= \sum_x xP(X=x) + \sum_y yP(Y=y) \\
&= \mathrm{E}[X] + \mathrm{E}[Y]
\end{aligned}$$

3. Further assume that $c \in$ is a scalar and is not a random variable, show that

$$\mathrm{E}[cX] = c\,\mathrm{E}[x]$$

$$\begin{aligned}
\mathrm{E}[cX] &= \sum_x cxP(X=x) \\
\mathrm{E}[cX] &= c\sum_x xP(X=x) \\
\mathrm{E}[cX] &= c\,\mathrm{E}[X]
\end{aligned}$$

4. Assuming X being a discrete random variable, show that:

$$Var(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$$

We know that $Var(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$,so

$$\begin{aligned}
Var(X) &= \mathrm{E}[(X - \mathrm{E}[X])^2] \\
&= \mathrm{E}[X^2 - 2X\,\mathrm{E}[X] + (\mathrm{E}[X])^2] \\
&= \mathrm{E}[X^2] - \mathrm{E}[2X\,\mathrm{E}[X]] + \mathrm{E}[(\mathrm{E}[X])^2] \\
&= \mathrm{E}[X^2] - 2(\mathrm{E}[X])^2 + (\mathrm{E}[X])^2 \\
&= \mathrm{E}[X^2] - (\mathrm{E}[X])^2
\end{aligned}$$