Assignment - Machine Learning

Snehanshu Saha
PES University

Instructions:

Solve any one problem below. The choice of the problem shall be determined by the instructor in class. Solution has to be typed using LaTeX rendering. Word documents and PDF generated from MS Word will NOT be accepted.

Problem Set

Gradient Learning and learning rates

The gradient descent update rule is given by

$$\mathbf{w} := \mathbf{w} - \alpha \cdot \nabla_{\mathbf{w}} f$$

where f is the loss function. When the learning rate, α, is too small, then convergence takes a long time. However, when the learning rate is too large, the solution diverges.

For a function f, the Lipschitz constant is given by $\max|\nabla_{\mathbf{w}} f|$. Therefore, by setting $\alpha = \frac{1}{L}$, we have $\Delta\mathbf{w} \leq 1$, constraining the change in the weights. This makes it optimal to set the learning rate to the reciprocal of the Lipschitz constant.

- Show that, for the Least Square Cost function, the Lipschitz constant is given by the right hand side of the inequality:

$$\boxed{\frac{\|g(\mathbf{w}) - g(\mathbf{v})\|}{\|\mathbf{w} - \mathbf{v}\|} \leq 2K \left\|\mathbf{X}\mathbf{X}^T\right\| - 2\left\|\mathbf{y}^T\mathbf{X}\right\|}$$

- Show that, for the binary cross entropy function, the Lipschitz constant is given by

$$\boxed{L = \frac{1}{2m}\|\mathbf{X}\|}$$

where m is the number of training examples.

- The softmax loss is defined as

$$g(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}[y^{(i)} = j]\log\left(\frac{e^{\mathbf{\Theta}_j^T\mathbf{x}^{(i)}}}{\sum_{l=1}^{k} e^{\mathbf{\Theta}_l^T\mathbf{x}^{(i)}}}\right)$$

  Show that the Lipschitz constant for the softmax loss function is given by

$$\boxed{L = \frac{k-1}{km}\|\mathbf{X}\|}$$

Activation Function

The activation function, SBAF is as follows:

$$y = \frac{1}{1 + kx^{\alpha}(1-x)^{1-\alpha}};$$

- Show that SBAF is not a probability density function!
- Show that SBAF can be mathematically linked to binary logistic regression.
- Show that SBAF satisfies Cybenko approximation theorem (refer to the classic 1989 paper by Cybenko).
- Show that SBAF is a solution to the first order differential equation:

$$f(x,y) = \frac{dy}{dx} = \frac{y(1-y)}{x(1-x)} * (\alpha - x)$$

- Show that the RHS of the above differential equation is Lipschitz continuous.

Bayesian learning

Minimum Error Classifier is given as

$$P(error) = \int P(error, X)dx$$
$$P(error) = \int P(error \mid X)p(X)dx$$
$$P(error) = \int min[P(\omega_1 \mid X), [P(\omega_2 \mid X)]p(X)dX$$

If P(ω_1 | X) = P(ω_2 | X), the classifier fails to arrive at a decision. Define the zero-one loss function as:

$$\lambda_{ij} = 1 \quad when \quad i \neq j$$
$$\lambda_{ij} = 0 \quad when \quad i = j$$

$1 \leq$ i $\leq$ c and $1 \leq$ j $\leq$ c where c is the total number of classes. The expressions for risk are as follows:

$$R(\alpha_i \mid X) = \sum_{j=1}^{c} P(\omega_j \mid X)$$
$$R(\alpha_i \mid X) = 1 - P(\omega_i \mid X)$$

Clearly, P(ω_i | X) has to be maximized in order to minimize R(α_i |X). The posterior probability corresponding to the class ω_i shall determine the outcome of the classifier in favor of that class.

- Show that, under zero-one loss function, he bound on risk under ***Min-Max Criterion*** is given as

$$R < \int_{R_1+R_2} (P(\omega_1)p(X \mid \omega_1) + (1 - P(\omega_1))p(X \mid \omega_2))dX$$

- Show that, $P(error)$, error of misclassification in the binary classification problem is bounded above by

$$P^{\beta}_{(\omega_1)} P^{1-\beta}_{(\omega_2)} \int p^{\beta}(x \mid \omega_1) p^{1-\beta}(x \mid \omega_2) dx, \qquad 0 \leq \beta \leq 1$$

  Note: We know,

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)p(x)dx$$