Machine Learning Worksheet 5

Linear Regression

1 Probability Theory

Problem 1. Let X have a continuous cdf $F_X(x)$. Define the random variable Y as $Y = F_X(X)$. Assuming that $F_X(x)$ is strictly increasing, how is Y distributed? Show your work.

2 Weighted Linear Regression

Consider a linear regression problem in which we want to "weight" different training examples differently. Specifically, suppose we want to minimize

$$E(\boldsymbol{w}) = \frac{1}{2}\sum^{N} \theta_n \left(z_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\right)^2$$

Problem 2. We already worked out what happens for the case where all the weights θ_n are the same. In this problem, we will generalize some of those ideas to the weighted setting, and also implement the locally weighted linear regression algorithm.

1. Show that $E(\boldsymbol{w})$ can also be written

$$E(\boldsymbol{w}) = (\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w})^T \boldsymbol{\Theta}(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}) \tag{1}$$

   for an appropriate diagonal matrix $\boldsymbol{\Theta}$, and where $\boldsymbol{\Phi}$ and $\boldsymbol{z}$ are as defined in class. State clearly what $\boldsymbol{\Theta}$ is.

2. Now let all the θ_n equal 1. By differentiating Eq. 1 with respect to $\mathbf{w}$, derive the normal equations for the least squares problem, as given in class.

3. Generalize the normal equations to the case of arbitrary θ_ns.

4. Suppose we have a training set $(\boldsymbol{x}_n, z_n)$; $n = 1, \ldots, N$ of N independent examples, but in which the z_n were observed with differing variances. Specifically, suppose that

$$p(z_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathcal{N}(z_n|\boldsymbol{w}^T\Phi(\boldsymbol{x}_n), \sigma_n^2)$$

   where the σ_n are fixed, known, constants. Show that finding the maximum likelihood estimate of $\boldsymbol{w}$ reduces to solving a weighted linear regression problem. State clearly what the θ_n are in terms of the σ_n.

3 Basisfunctions

Problem 3. Show that the tanh function and the logistic sigmoid function are related by

$$\tanh(x) = 2\sigma(2x) - 1$$

Thus, show that a general linear combination of logistic sigmoid functions of the form

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\left(\frac{x - \mu_j}{s}\right)$$

is equivalent to a linear combination of tanh functions of the form

$$y(x, \boldsymbol{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\left(\frac{x - \mu_j}{2s}\right)$$

and find expressions to relate the new parameters $\{u_0, \ldots, u_M\}$ to the original parameters $\{w_0, \ldots, w_M\}$.

4 Computation

Problem 4. X is a random variable with $X \sim U[0,1]$ (i.e. uniformly distributed in the interval from 0 to 1). Furthermore

$$f(X) = \frac{\sin(12(X + 0.2))}{X + 0.2}$$

and $Y = f(X) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0,1)$ (i.e. gaussian noise with mean 0 and variance 1). Sample $N_1 = 10$ and $N_2 = 100$ many pairs from $(X, f(X))$ and fit polynomials of degree 3 *and* degree 9. Prepare one plot for N_1 and one plot for N_2, showing the resulting polynomials. Furthermore, plot the results you get using an $\ell 2$ penalty for the polynomial of degree 9 with penalty $\lambda = 5$.

5 Ridge regression

Problem 5. Show that the following holds: The ridge regression estimates can be obtained by ordinary least squares regression on an augmented dataset: Augment the design matrix $\boldsymbol{\Phi}$ with p additional rows $\sqrt{\lambda}\boldsymbol{I}$ and augment $\boldsymbol{y}$ with p zeros.

Problem 6. Using singular value decomposition of the design matrix $\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ show that the output on the training set fitted with the ridge regression solution $\hat{\boldsymbol{w}}^{ridge}$ can be written as

$$\sum_j \left(\frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j \boldsymbol{u}_j^T\right) \boldsymbol{y}$$

where $\boldsymbol{u}_j$ are the columns of $\boldsymbol{U}$, d_j the elements of $\boldsymbol{D}$ and λ the cost factor of the $l2$ regularization. What is the interpretation of this formula?

6 Multi-output linear regression

Problem 7. In class, we only considered functions of the form $f : \mathbb{R}^n \to \mathbb{R}$. What about the general case of $f : \mathbb{R}^n \to \mathbb{R}^m$? For linear regression with multiple outputs, write down the loglikelihood formulation and derive the MLE of the parameters.