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.

Problem 2. Show that the sum of two independent Gaussian random variables ($\boldsymbol{X_1}$ and $\boldsymbol{X_2}$) is Gaussian. Some of the properties of Gaussians mentioned in the lecture can help.

Problem 3. Let $Z = (X, Y)$ be a bivariate normal distributed random variable. Furthermore, let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Assume that $\rho(X, Y) = 0$. Show that in this case X and Y are independent.

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 4. 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

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

   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_n.

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.

5. With *ordinary* linear regression it may be a good idea to *rescale* the *columns* of the design matrix – in particular when using nonlinear basis function expansions (e.g. like polynomial expansion). Using the normal equations, prove that rescaling the design matrix does not change the predicted values for some test dataset.

3 Basisfunctions

Problem 5. 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 Ridge regression

Problem 6. 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 $\boldsymbol{z}$ with p zeros.

Problem 7. 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 $\ell 2$ regularization. What is the interpretation of this formula?

5 Multi-output linear regression

Problem 8. 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.

6 Bayesian Linear Regression

Problem 9. $\star$ We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases (see tower equalities). Prove the following matrix identity

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}$$

and, using it, show that the uncertainty $\sigma_N^2(\boldsymbol{x})$ associated with the bayesian linear regression function given by eq. (26) on the slides satisfies

$$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x})$$