Tutorial 2

CS 337 Artificial Intelligence & Machine Learning, Autumn 2021

September 2021

Problem 1. Consider a data set in which each data point y_i is associated with a weighting factor r_i, so that the sum-square error function becomes

$$\frac{1}{2}\sum_{i=1}^{m} r_i(y_i - w^T\phi(x_i))^2$$

Find an expression for the solution w^* that minimizes this error function. The weights r_i's are known before hand. (Exercise 3.3 of Pattern Recognition and Machine Learning, Christopher Bishop).

Problem 2. Equivalence between Ridge Regression and Bayesian Linear Regression (with fixed σ^2 and λ): Consider the Bayesian Linear Regression Model

$$y = \mathbf{w}^T\phi(\mathbf{x}) + \varepsilon \; and \; \varepsilon \sim \mathcal{N}(0, \sigma^2)$$
$$\mathbf{w} \sim \mathcal{N}(0, \alpha I) \; and \; \mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\mu_m, \Sigma_m)$$
$$\mu_m = (\lambda\sigma^2 I + \phi^T\phi)^{-1}\phi^T\mathbf{y} \; and \; \Sigma_m^{-1} = \lambda I + \phi^T\phi/\sigma^2$$

Show that $\mathbf{w}_{MAP} = \underset{\mathbf{w}}{\text{argmax}} \; \Pr(\mathbf{w} \mid \mathcal{D})$ is the same as that of *Regularized Ridge Regression.*

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda\sigma^2||\mathbf{w}||_2^2$$

In other words, The Bayes and MAP estimates for Linear Regression coincide with that of *Regularized Ridge Regression.*

Problem 3. Ridge Regression and Error Minimization:

1. *Prove the following Claim:*
   The sum of squares error on training data using the weights obtained after minimizing ridge regression objective is greater than or equal to the sum of squares error on training data using the weights obtained after minimizing the ordinary least squares (OLS) objective.

   More specifically, if ϕ and $\mathbf{y}$ are defined on the training set $\mathcal{D} = \{(\mathbf{x}_1, y_1)...(\mathbf{x}_m, y_m)\}$ as

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & ..... & \phi_n(\mathbf{x_1}) \\ . & & & \\ . & & & \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & ..... & \phi_n(\mathbf{x}_m) \end{bmatrix} \tag{1}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_m \end{bmatrix} \tag{2}$$

   and if

$$\mathbf{w}_{Ridge} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$$

   and

$$\mathbf{w}_{OLS} = \underset{\mathbf{w}}{\text{argmin}} \; ||\phi\mathbf{w} - \mathbf{y}||_2^2$$

   then you should prove that

$$||\phi\mathbf{w}_{Ridge} - \mathbf{y}||_2^2 \geq ||\phi\mathbf{w}_{OLS} - \mathbf{y}||_2^2$$

2. If it is the case that ridge regression leads to greater error than ordinary least squares regression, then why should one be interested in ridge regression at all?

Problem 4. Gradient descent is a very helpful algorithm. But it is not guaranteed to converge to global minima always. Give an example of a continuous function and initial point for which gradient descent converges to a value which is not global minima.

Problem 5. In class, we have illustrated Bayesian estimation for the parameter μ of a Normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, assuming that σ was known by imposing a Normal (conjugate) prior on μ. Now suppose that the parameter μ is known and we wish to estimate σ^2. What will be the form of the conjugate prior for this estimation procedure? If $\mathcal{D} = X_1, X_2, X_3, ..., X_n$ is a set of independent samples from this distribution, after imposing the conjugate prior, compute the form of the likelihood function $\mathcal{L}(\theta)$, the posterior density $P(\theta \mid \mathcal{D})$ and the posterior probability $P(X \mid \mathcal{D})$. Again, you can ignore normalization factors.

Problem 6. Consider a linear model of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$ (i.e. $\mathbb{E}[\epsilon_i \epsilon_j] = \sigma^2$ when $i = j$), show that minimizing E_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer. (Problem 3.4 from Bishop, PRML)