Homework 2

Aukkawut Ammartayakun
CS 539 Machine Learning

Spring 2023

Problem 1

Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\right)^2 \tag{1}$$

Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points

Problem 2

We showed in the class that the conjugate prior for a Gaussian distribution with unknown mean and unknown precision (inverse variance) is a normal-gamma distribution. This property also holds for the case of the conditional Gaussian distribution $p(t|x, w, \beta)$ of the linear regression model. If we consider the likelihood function (3.10),

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}) \tag{2}$$

then the conjugate prior for $\mathbf{w}$ and β is given by

$$p(\mathbf{w}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \beta^{-1}S_0)\mathrm{Gamma}(\beta|a_0, b_0) \tag{3}$$

Show that the corresponding posterior distribution takes the same functional form, so that

$$p(\mathbf{w}, \beta|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}S_N)\mathrm{Gamma}(\beta|a_N, b_N) \tag{4}$$

and find expressions for the posterior parameters $\mathbf{m}_N$, S_N, a_N, and b_N.

Problem 3

Show that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\mathbf{w}$ whose decision boundary $\mathbf{w}^\top \phi(\mathbf{x}) = 0$ separates the classes and then taking the magnitude of $\mathbf{w}$ to infinity

Problem 4

Show that the Hessian matrix $\mathbf{H}$ for the logistic regression model, given by (4.97),

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^\top = \mathbf{\Phi}^\top \mathbf{R}\mathbf{\Phi} \tag{5}$$

is positive definite. Here $\mathbf{R}$ is a diagonal matrix with elements $y_n(1-y_n)$, and y_n is the output of the logistic regression model for input vector $\mathbf{x}_n$. Hence show that the error function is a concave function of $\mathbf{w}$ and that it has a unique minimum.

Problem 5 (Likelihood Estimate for Gamma regression)

Gamma distribution is defined by

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{6}$$

1. Write down the probability in general form of the exponential distribution family, and find natural parameter η, $u(x)$, $h(x)$, and $g(\eta)$.
2. Let's assume we have a set of data points (t_i, x_i) $i = 1, \ldots, N$, and we assume t_i follows a Gamma distribution where its mean is defined by

$$y_k = \exp\left(w_0 + w_1 x_k\right) \tag{7}$$

and the conditional distribution is

$$f(t_k|y_k) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu t_k}{y_k}\right)^\nu \frac{1}{y_k} e^{-\frac{\nu t_k}{y_k}} \tag{8}$$

discuss how you can find maximum likelihood estimates of w_0 and w_1 using a gradient ascent algorithm. Derive the gradient and discuss whether the likelihood function is a concave function of the w_0 and w_1 or not.

Problem 6 (Laplacian Prior)

Laplacian prior for the weights of a linear (or logistic) regression will turn into Lasso regularization. Laplacian distribution on w is defined by

$$p(\mathbf{w}) = \frac{1}{2b} \exp\left(-\frac{|\mathbf{w}|}{b}\right) \tag{9}$$

which can be defined for weights of the model (except the intercept), where we assume different weigths are independent. b is a hyperparameter.

1. Let's assume we have $D = \{(\mathbf{x}_i, t_i) | i = 1, \ldots, N\}$ and we want to build a linear regression model with the Laplacian prior on the model weights. Define the likelihood and prior term here, and show it turns to a lasso regression. You can assume weights share the same hyperparameter.
2. Lasso regression is defined by

$$E_D(\mathbf{w}) = -\frac{1}{2} \sum_{i=1}^{N} (t_i - \mathbf{w}^\top \phi(\mathbf{x}_i))^2 - \lambda \sum_{j=1}^{M} |w_j| \tag{10}$$

We can use a gradient descent algorithm to find the model parameters, but the issue is that derivative of $|\mathbf{w}|$ has a discontinuity at zero. A remedy is to rewrite the optimization by

$$E_D(\mathbf{w}) = -\frac{1}{2} \sum_{i=1}^{N} (t_i - \mathbf{w}^\top \phi(\mathbf{x}_i))^2 - \lambda \sum_{j=1}^{M} \frac{w_j^2}{|w_j|} \tag{11}$$

where, you replace the term in denominator of the regularization term by a known value. Let's assume, you are in the r^{th} iteration of a gradient descent algorithm (r represents the iteration), and your partial derivative for j^{th} weight is defined by

$$\frac{\partial E_D^{(r)}(\mathbf{w})}{\partial w_j} \approx \sum_{i=1}^{N} \phi(\mathbf{x}_i)\left(t_i - {\mathbf{w}^{(r)}}^\top \phi(\mathbf{x}_i)\right) - \lambda \frac{w_j^{(r)}}{\max\left\{\epsilon, |w_j^{(r-1)}|\right\}} \tag{12}$$

where, ϵ has a small vale, like 0.0001. Complete the update rule for all other weights in the model and show its result in a simulated data.

3. Create 100 sample data points for $t_i = 1 + 0.001\mathbf{x}_i - 2\mathbf{x}_i^2 + \epsilon_k$ where ϵ_k has normal distribution with a mean zero and variance of 0.1. Show how the estimated weights will change as a function of λ. For x, you can draw 100 random values from a normal distribution with mean 0 and variance 1. To find the model parameters you can use the gradient descent algorithm we discussed here.