Machine Learning November 4, 2020
Sharif University of Technology CE 40-712
Dr. A. Hosseini

Due Date: 1399.08.27

Homework 4

Closed Form Solution For Regression

- For a given data set, let the least-squares cost function be written as $||y - Xw||^2$, where X is a matrix with one data point per row, y is a vector with one response value per row, show the closed from solution for above equation is $w^* = (X^\intercal X)^{-1} X^\intercal y$
- Now show the closed from solution for ridge regression is $w^* = (X^\intercal X + \lambda I)^{-1} X^\intercal y$

Regression Shrinkage Methods

- Show that linear regression with L_2 regularization can be converted to least squares linear regression by adding some data points.
- Is there any senario in which L_1 regularization fails?
- Consider the following loss function :

  $$L(w, \lambda_1, \lambda_2) = |y - Xw|^2 + \lambda_1 ||w||_2^2 + \lambda_2 ||w||_1$$

  Show that this loss function is equivalent to L_1 regularized loss function by adding some data points.

Regression and Gradient Descent

Suppose you have following model :

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + \epsilon \quad where\, \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Write down an expression for $P(y|x_1, x_2)$
- Assume you are given a set of training observations $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$ for $i = 1, 2, ..., n$ write down the conditional log likelihood of this training data. Drop any constants that do not depend on the parameters $w_0, ..., w_4$
- Write down a function $f(w_0, w_1, w_2, w_3)$ that can be minimized to find the desired parameter estimates.
- Calculate the gradient of $f(w)$ with respect to the parameter vector w.
- Write down a gradient descent update rule for w in terms of $\nabla_w f(w)$

Active regression In active regression, we have the model as a black box and it's possible to get output for arbitary input. Consider linear regressor with base functions $\{\phi_i(x)_{i=1,...,n}\}$ and $f(x;w) = \sum_{i=1}^{m} \phi_i(x)w_i$. If $\phi(x) = (\phi_1(x), ..., \phi_m(x))^T$ and $X = (\phi(x_1), ..., \phi(x_n))^T$ then we have $y = Xw + e$ where $e \sim N(0, I)$.

- If we get the output for the same input multiple times , prove that :

  $$\hat{w} \sim N(w^*, (X^T X)^{-1})$$

  where W^* are true parameters.

- Calculate the variance of estimation.

- Now suppose that model's input is sampled from $Q(x)$ and the goal is to minimize estimation variance. So we can define the loss function as

  $$J(X) = E_{x \sim Q}\{Var(\hat{y}(x))\}$$

  . Suppose x is one-dimensional and in the range $[-1, 1]$ and $\phi(x) = (1, x)^T$.

  - If we define $A_n = X^T X$, prove

    $$A_n = \sum_{i=1}^{n} \phi(x_i)\phi(x_i)^T$$

    and is symmetric positive definite. Write A_n as a function of n and x_i.

  - If $Q(x)$ is a symmetric distribution with mean zero and variance v^2, calculate the loss function.

Probabilistic Modeling of Regression Suppose for each input x we are given a target output y. In probabilistic approach we have $p(y|x; w, \beta) = f(x, w) + \epsilon$ which $\epsilon \sim \mathcal{N}(0, \beta^{-1})$ Here, $f(x, w)$ is a nonlinear function predicting the output value of an input x.

1. Show that maximizing Log-Likelihood in this context is the same as minimizing Sum Square Error.

2. Now assume that we know $p(w; \alpha) = N(0, \alpha^{-1}I)$ as a prior. Using the Bayes' theorem and MAP approach, find the maximum of the posterior. Show that this result is the same as the optimal value of Ridge-Regularized Sum Square Error.

3. Now assume that we are given another prior knowledge about parameters w. Here, we know that $p(w; \alpha) = Laplace(0, \alpha^{-1}I)$. Show that with this prior we get the optimal value

   of Lasso-Regularized Sum Square Error.

Bayesian linear regression

- Suppose prior distribution is a multivariate gaussian $w \sim N(M_0, S_0)$ with likelihood $y \sim N(w^T x, \beta^{-1})$. Prove that postrior distribution is $w \sim N(m_N, S_N)$ with

  $$m_N = S_N(S_0^{-1} m_0 + \beta X^T Y), \quad S_N^{-1} = S_0^{-1} + \beta X^T X$$

- If prior distribution is the special case of $w \sim N(0, \alpha^{-1}I)$ calculate posterior.
- For the special case of diagonal covariance matrix, we want to prove posterior distribution is equal by observing whole data or observing data points one by one (at each point, previous posterior is seen as the prior for the next stage). First prove that

$$S_{N+1}^{-1} = S_N^{-1} + \beta x_{n+1} x_{n+1}^T$$

$$m_{N+1} = S_{N+1}(S_N^{-1} m_N + \beta x_{n+1} y_{n+1})$$

  and using these recursive equations prove what we want.

- By using the following lemma, prove increasing number of data points decreases variance of predictive distribution.

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

Practical In this problem, we are trying to train a model to estimate car price based on its features. All you need is provided at file ”Toyota.xls”. The first sheet defines the features and second sheet is the data itself. Divide dataset into 70%, 15% and another %15 for train, validation and test. Dont́ use ”ID” and ”model” features and packages like ”sklearn” (numpy and pandas are OK). Do not forget to prepare a report!

- Train a simple regressor using SGD and report MSE for test dataset. Plot MSE per iteration for validation and test datasets.
- Now train a ridge regressor ($\lambda = 1$) and report accuracy for validation and test datasets.
- Repeat part 1 for $||Y - X\omega||_1$
- Compare regressors of part 1 and 3 by investigating weights of the model. Also try to set λ to zero at part 2 and report the results and talk about advantages of ridge regression.