STAT561 Homework 2

(Suggested) Due Date: February 13, 2019

1 Analysis

1.1 Logistic Regression

Recall that under the logistic model, with $y \in \{-1, 1\}$, the negative log likelihood is

$$\begin{aligned} L &= -\ln p(Y|X, \beta) \\ &= -\sum_{i=1}^{n} \ln \sigma(-y_i \beta^T x_i) \end{aligned}$$

where σ is the logistic function. Since L is convex in β, we can use gradient descent to find the β achieving the global minimum of L.

1. Derive the gradient $\nabla_\beta L$. It will be useful to first find $\sigma'(z)$.

2. The parameter update in gradient descent takes the form of $\beta_{k+1} = \beta_k - \nabla_\beta L$. Give a geometric interpretation of this update, taking into account the particular form the gradient takes (i.e. an explanation beyond the fact that the update moves "down" the loss function").

3. How does the gradient change with the addition of l_2 regularization? How about with l_1 regularization?

4. Show that in the univariate case (where β is a scalar), L is convex.

5. *Optional:* In the multivariate case, L is convex when its Hessian is positive semidefinite. Show that L is convex.

1.2 Gaussian Process Regression

For GP regression, we assume our data follows the model

$$y_i = f(x_i) + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

and we place a prior on f in the form

$$p(f|\mu(\cdot), K(\cdot,\cdot)) = GP(\mu(\cdot), K(\cdot,\cdot))$$

where μ is the mean function and K is the covariance kernel. With the goal of obtaining a predictive distribution $p(Y^*|X^*, Y, X, \mu, K, \sigma^2)$, we will specify μ, K, and σ^2.

1. Is $K(x_i, x_j) = x_i^T x_j$ a valid kernel?

2. Is $K(x_i, x_j) = x_i^T x_j - 1$ a valid kernel?

3. Is $K(x_i, x_j) = \frac{2}{5}||x_i||_2^2 + \frac{3}{5}||x_j||_2^2$ a valid kernel?

4. The figure below shows confidence intervals from the posterior over f. What does it look like σ^2 was set to?

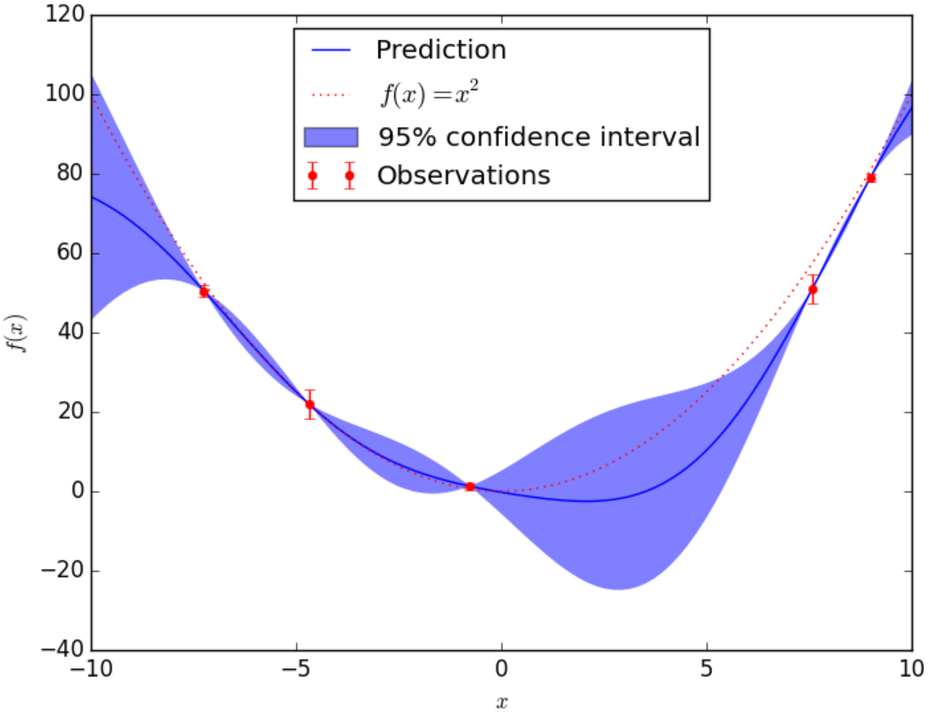

Figure 1: Figure from `https://scikit-learn.org/0.17/auto_examples/gaussian_process/plot_gp_regression.html`

2 Computation

2.1 Logistic Regression

The state of North Carolina maintains a public voter record which includes personal information and voting histories for each registered voter. The dataset `hw2_voting.csv` made available on the course website is a sample of $1,000$ registered voters in North Carolina, and includes each individual's race and ethnicity as recorded from their voter registration form, the party they registered as a member of, their gender and age, an indicator for if they registered with a driver's license, their date of registration, and their voting history in major statewide elections from 2008 through 2016. Voting history includes only whether or not someone cast a vote in an election. It does not include who they voted for.

You may use auxiliary software packages for each of the following except 2.

1. Randomly subset 10% of the data as a hold-out set for model validation.

2. Model the probability of voting in the 2016 general election by fitting a logistic regression using gradient descent. Include age as a predictor; include other features as you'd like. Transform the features as needed to obtain a better fit.

3. Validate your model by checking the accuracy on the held out data. Why should you obtain at least 70% accuracy?

4. Plot the ROC curves for the training set and the validation set and check each AUC.

5. Plot the predicted probability of voting in the 2016 election as a function of age, with all other variables held constant (if you included others). What's the relationship between age and voting?

6. Try adding l_1 and then l_2 regularization to the loss function. With $\lambda = 1$, does your validation performance improve?

2.2 Gaussian Process Regression

The `hw2_geyser.csv` data available on the course website is a set of measurements of geyser eruptions in Yellowstone National Park, specifically from the geyser named Old Faithful. Each row includes the length of an eruption and the time until the next eruption, both measured in minutes. Here we will predict time until next eruption as a function of eruption length.

1. Compute the covariance matrix $K(\mathbf{X}, \mathbf{X})$ of the data using the Gaussian kernel,

$$k(x_i, x_j) = \nu \exp(\frac{-||x_i - x_j||^2}{2\tau^2})$$

   with $\nu = 1$ and $\tau = 1$.

2. Let $\mathbf{X}^*$ range from 1 to 6 and consist of 100 evenly spaced points. Find the predictive distribution $P(\mathbf{Y}^*|\mathbf{X}^*, \mathbf{X}, \mathbf{X})$.

3. With $\sigma^2 = 0$, sample four draws of $\mathbf{Y}^*$ from the predictive distribution and plot them in separate colors. Note that when $\sigma^2 = 0$, this is equivalent to sampling four draws of $f(\mathbf{X}^*)$, so you will be plotting each of the four functions evaluated at $\mathbf{X}^*$.

4. Plot an estimate of $E(\mathbf{Y}^*)$, as well as the 95% confidence intervals for $\mathbf{Y}^*$ with $\sigma^2 = 1$. To do so, draw 1000 samples of $\mathbf{Y}^*$ and for each x_i^*, find the mean and 2.5^{th} and 97.5^{th} percentiles of each corresponding 1000 y_i^* values.

5. Normally, we'd validate the model by checking error on a validation set, as well as checking that the confidence intervals cover data appropriately often. Here, simply describe how well the model fits the data based on the plot from question 4.

6. You can tune multiple hyperparameters like ν and τ by checking the model error for models fit with a variety of combinations of the hyperparameters, a process known as grid search. Since you aren't validating the model with a specific error metric, try adjusting the hyperparameters and assess how this influences the model fit.

7. Remove data points from $\mathbf{X}$ with eruption lengths between 2.75 and 3.25, then repeat the exercises with this data. How do the confidence intervals differ betweeen the datasets?