DS102 - Homework 3

If you are submitting a handwritten version please make sure your answers are legible as you may lose points otherwise.
Data science is a collaborative activity. While you may talk with others about the homeworks, we ask that you write your solutions individually. If you do discuss the homework with others please include their name in your submission.
Due by: 1:59pm, Tuesday 8th October, 2019

1. (20 points) Here we will consider MAP estimation of regression weights under the linear Gaussian model.

   Specifically, suppose that $x = (x_1, \ldots, x_n)$ are fixed d-dimensional vectors, and suppose we obtain observations:

   $$y_i = \beta^\top x_i + \epsilon_i, i \in \{1, \ldots, n\},$$

   where as usual $\epsilon_i \sim N(0, \sigma^2)$ are independent and $\beta \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}^+$ are unknown.

   We denote $y = (y_1, \ldots, y_n)$, and X the matrix with i^{th} row equal to x_i. Here we will model the regression weights as a random variable, and place the following prior distribution on them:

   $$\beta \sim N(0, \sigma_\beta^2 \cdot I).$$

   In words, this means that every entry of β is distributed as $N(0, \sigma_\beta^2)$, and that the entries are independent.

   (a) (7 points) Write the posterior distribution for β after observing the data, $p(\beta|X, y)$. It's fine to leave it in terms of a proportionality constant.

   (b) (7 points) Show that the MAP estimator of β,

   $$\hat{\beta}_{MAP} := \arg\max_\beta p(\beta|X, y)$$

   is also the minimizer of the regularized least squares equation,

   $$\arg\min_\beta \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2$$

   for some non-negative $\lambda \in \mathbb{R}$.

   (c) (6 points) (i) What is λ, as a function of the parameters in this problem?
   (ii) In the regularized least squares problem, λ is a regularization term, where large λ penalizes solutions with large norm of the regression weight vector. Use the form of λ from part (i) to discuss how our modelling choices (i.e. choice of σ_β^2) influence this regularization.

2. (20 points) For this exercise make sure to attach both plots and code with your solutions.

   In this question we compare the effectiveness of different linear classifiers and regression models on the Boston housing and Iris dataset. Use the provided code at this link for this question.

   The Boston housing dataset involves predicting average housing prices in an area given features of the area (e.g. average number of rooms per house, average age of owners, etc...). The Iris dataset involves predicting between 3 types of irises given the features of the plant (e.g. petal length, petal width, etc...), however here we only predict on the first two iris types to keep the dataset binary.

   (a) (2 points) Fill in the `regression_predict` and `regression_least_squares` functions.

   (b) (2 points) Fill in the `logistic_predict` and `logistic_cross_entropy_loss` functions.

   (c) (2 points) Fill in the `gradient_descent` function.

   (d) (6 points) For the Boston housing dataset we will use the root mean squared error (RMSE) as our error metric:

$$RMSE(\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} = \frac{1}{\sqrt{n}}\|\hat{y} - y\|_2,$$

   where n is the number of datapoints, $\hat{y}$ is our predicted price vector and y is the true price vector.

   Train a linear regression model by computing

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|X_{train}\beta - y_{train}\|_2^2 + \lambda\|\beta\|_2^2.$$

   And then test the model by computing $RMSE(X_{test}\hat{\beta})$.

   - Plot the RMSE by using `boston_X_train` as X_{train} and `boston_X_test` as X_{test} use regularization values $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. This corresponds to using the raw numerical features without any modification.
   - Plot the RMSE by using `boston_poly_X_train` as X_{train} and `boston_poly_X_test` as X_{test} use regularization values $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. This corresponds to using features

$$\phi(x) = \left[1, x_1^2, x_1x_2, x_2^2, \ldots, x_1x_2^4, x_2^5\right]^\top$$

     where x corresponds to the original features.

   (e) (2 points) Give an interpretation to your results in part (d). Did the value of the regularization term λ matter for both featurizations? Why or why not? Which featurization performed better? Why?

(f) (6 points) For the Iris dataset we will use the mean average error (MAE) as our metric:

$$MAE(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\hat{y}_i \neq y_i)$$

where $\hat{y}$ is our vector of $0-1$ predictions and y is the true vector of labels.

Train your logistic regression model using gradient descent on the iris training set. Try to get the lowest error rate on the iris test set by tuning the learning rate and number of iterations. Note that your logistic model outputs a real number $p \in [0, 1]$ which you need to convert to a $0-1$ decision. To do so use the following decision function:

$$\delta(p) = \mathbb{1}(p \geq 0.5).$$

What is the best MAE that you achieved?

3. (20 points) The goal of this question is to get a better understanding of the Poisson distribution and conjugate priors. Assume that we are recording the number of cars crossing an intersection per day for a traffic survey. We have a dataset that consists of the number of cars that have crossed the intersection collected over a period of n days:

$$\{k_1, k_2, \ldots, k_n\}.$$

A good way to model these counts if we assume they are iid samples from a Poisson distribution

$$\mathbb{P}(k_i = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

where $\lambda > 0$ is an unknown parameter we wish to determine.

(a) (3 points) Given a random variable $K \sim Pois(\lambda)$ compute $\mathbb{E}[K]$.
(Hint: recall that $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, for any $x \in \mathbb{R}$.)

(b) (2 points) Compute the likelihood of the data $\mathbb{P}(k_1, k_2, \ldots, k_n|\lambda)$.

(c) (2 points) Find the maximum likelihood estimator of λ with respect to the collected data.

(d) (1 point) What is the relationship between the MLE and the expected value computed in part (a)?

(e) (5 points) Now assume we put a prior distribution of Gamma on the parameter λ

$$\lambda \sim Gamma(\alpha, \beta),$$

for $\alpha > 0$ and $\beta > 0$ then the pdf of λ is given by

$$\mathbb{P}(\lambda|\alpha, \beta) = \frac{\beta^\alpha \lambda^{\alpha-1} \exp(-\beta\lambda)}{\Gamma(\alpha)},$$

where

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx.$$

Show that the posterior distribution $\mathcal{P}(\lambda|k_1, k_2, \ldots, k_n)$ is a Gamma distribution. (Hint: As in discussion 4 you do not need to compute any normalization factors to show that the distribution is a Gamma.)

(f) (5 points) Compute the maximum a posteriori (MAP) estimate of λ,

$$\operatorname*{argmax}_{\lambda} \mathcal{P}(\lambda|k_1, k_2, \ldots, k_n).$$

(g) (2 points) Give an interpretation of the posterior and prior Gamma distributions and an interpretation of the MAP estimate.