Problem Set #2

ECE 4424 / CS 4824 - Machine learning
Virginia Tech

September 22, 2020

- Feel free to collaborate with other classmates in doing the homework. Please indicate your collaborators with their student ID. You should, however, write down your solution yourself. Please try to keep the answers brief and clear.
- Whenever you need clarification, please post the related questions on Piazza under the corresponding homework folder.
- Total: 100 points
- **Due date: 10/5/2020, 11:59PM ET**
- Late submission: each student will have a total of **four** free late (calendar) days to use for homeworks. Once these late days are exhausted, any assignments turned in late will be penalized 20% per late day. However, no assignment will be accepted more than three days after its due date. Each 24 hours or part thereof that a homework is late uses up one full late day.

1 Maximum likelihood estimation

Question 1.1 (30 points): Suppose we throw a coin n times. Let $\mathcal{D} = \{y_1, ..., y_n\}$ denote the dataset we obtain, where $y_i \in \{0, 1\}$ is the outcome of the i-th throw (i.e., $y_i = 1$ if the coin comes up to be a head and 0 if it comes up to be a tail). We are interested in finding the maximum likelihood estimation of the probability that the coin comes up with a head, $\theta \in [0, 1]$, given this dataset $\mathcal{D}$.

a) (10 pts) Write out $P_\theta(\mathcal{D})$, i.e., the probability of observing the dataset $\mathcal{D}$ under the probability distribution is parameterized by θ. Hint: this probability distribution is NOT binomial distribution, since we assume that you have already observed the outcome of every coin.

b) (10 pts) Write out the log-likelihood, $\log P_\theta(\mathcal{D})$

c) (10 pts) Obtain the maximum likelihood estimation of θ given the dataset $\mathcal{D}$

2 Gaussian discriminant analysis

Question 2.1 (30 points): Suppose we are given a dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. We will model the joint distribution of (x, y) according to:

$$\begin{aligned} p(y) &= \phi^y (1-\phi)^{1-y} \\ p(x|y=0) &= \frac{1}{(2\pi)^{d/2}|\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^\top \Gamma^{-1}(x-\mu_0)\right) \\ p(x|y=1) &= \frac{1}{(2\pi)^{d/2}|\Gamma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^\top \Gamma^{-1}(x-\mu_1)\right) \end{aligned}$$

Here, the parameters of our model are ϕ, Γ, μ_0 and μ_1. Note that while there are two different mean vectors μ_0 and μ_1, there is only one covariance matrix Γ.

a) (5 pts) Suppose we have already fit ϕ, Γ, μ_0 and μ_1, and now want to make a prediction at some new query point x. Show that the posterior distribution of the label at x can be written as

$$p(y=1|x) = \frac{1}{1+\exp(-\theta^\top x + \theta_0)},$$

where the vector θ and scalar θ_0 are some appropriate functions of ϕ, Γ, μ_0 and μ_1 that you need to specify.

b) (25 pts) For this part of the problem only, you may assume d (the dimension of x) is 1, so that $\Gamma = \sigma^2$ is just a real number, and likewise the determinant of Γ is given by $|\Gamma| = \sigma^2$. Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\begin{aligned} \phi &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i = 1\} \\ \mu_0 &= \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = 0\} x_i}{\sum_{i=1}^{n} \mathbb{1}\{y_i = 0\}} \\ \mu_1 &= \frac{\sum_{i=1}^{n} \mathbb{1}\{y_i = 1\} x_i}{\sum_{i=1}^{n} \mathbb{1}\{y_i = 1\}} \\ \Gamma &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^\top \end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator function we have seen in class. The log-likelihood of the data is

$$\log P_\Theta(\mathcal{D}) = \log \Pi_{i=1}^{n} p(x_i, y_i) = \log \Pi_{i=1}^{n} p(x_i|y_i)p(y_i).$$

By maximizing $\log P_\Theta(\mathcal{D})$ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, Γ, μ_0 and μ_1 are indeed as given in the formulas above. (You may assume that there is at least one positive and one negative example, so that the denominators in the definitions of μ_0 and μ_1 above are non-zero.)

3 Programming assignment: Linear Regression (40 pts)

For the following programming assignment, please download the datasets and iPython notebooks from Canvas and submit the following:

- Completed and ready-to-run iPython notebooks. Note: we will inspect the code and run your notebook if needed. If we cannot run any section of your notebook, you will not receive any points for the task related to that section.
- Responses (texts, codes, and/or figures) to the following problems/tasks

In this programming exercise, you will build a linear regression model and apply it to a covid-19 sample dataset.

Task P1 (6 pts): Complete the codes that generate the three visualization graphs that show the trend of the epidemic progression ("People_tested", "Deaths", and "New_positive_cases"). Copy them to the solution file.

Task P2 (4 pts): Complete the function `predict_output`. Copy the the outputs of the code to the solution file.

Task P3 (6 pts): Let the regression cost function be given by

$$L_D(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w \cdot x_i)^2,$$

where $x_i \in \mathrm{R}^d$ is the input feature of dimension d, $y_i \in \mathrm{R}$ is the output response, and $w \in \mathrm{R}^d$ is the regression weights. Complete the function `weight_derivative` to calculate the derivative of the cost function with respect to regression weights w, i.e., $\frac{\partial}{\partial w} L_D(w)$. Note that this should be a d dimensional vector. Also copy the output of the code for the test example to the solution file.

Task P4 (5 pts): Complete the code section to perform the gradient decent in the function **regression_gradient_descent**. Copy the code to the solution file.

Task P5 (3 pts): Specify the initial weights, step size and tolerance for the function **regression_gradient_descent**. Print the outputs of the code.

Task P6 (3 pts): Use the learned weights to predict 'People_tested' in the last three weeks in the dataset. Copy the predictions to the solution file, and calculate the test error

$$\frac{1}{n_{\mathrm{tst}}}\sum_{i=1}^{n_{\mathrm{tst}}}(y_i^{\mathrm{tst}} - \hat{y}_i^{\mathrm{tst}})^2,$$

where n_{tst} is the number of test data, y_i^{tst} is the true label, $\hat{y}_i^{\mathrm{tst}}$ is the predicted label.

Task P7 (3 pts): Specify the initial weights, step size and tolerance for the function **regression_gradient_descent**. Print the outputs of the code.

Task P8 (4 pts): Use the learned weights to predict 'People_tested' in the last three weeks in the dataset. Find the value of the model predictions on the 10th day of the forecasting period. Also print the actual number of people tested on that particular day. Copy the predictions to the solution file, and calculate the test error. Note: here we are asking you to report the number before normalization. So you need to convert the prediction back to the unit of people.

Task P9 (6 pts): Explore on your own. Report your question of investigation, as well as your results/interpretation in the solution file.