Statistical inference and machine learning
Homework 1

- This assignment can be solved in groups of 1 up to 5 students. You must mention the name of all the participants. Note that all the students in a group will get the same grade.
- Deadline: Monday 26 October 2020, 23:59 (No late submissions will be accepted)
- Upload a single zip file on Moodle containing your solution and code. You can use any programming language.

1 Logistic Regression [70 pts]

An environmental scientist studying the forrest in the region of the Amazon has collected data through remote sensing. She is specifically interested in studying the proportion of land mass that is occupied by indigenous plants as compared to invasive species. She collects data $\{y_i, x_{i1}, x_{i2}, \ldots, x_{id}\}_{i=1}^n$, where the y_i are binary variables indicating whether a sampled region has either indigenous or invasive species. The $(x_{i1}, x_{i2}, \ldots, x_{id})$ are variables which record different physical properties of the region observed by the remote sensing satellite, where the sample is taken.

She also proceeds to collect new data through remote sensing from other regions of the Amazon where only measurements on the variables $(x_{i1}, x_{i2}, \ldots, x_{id})$ are taken. She wishes to predict the vegetation in each of these new regions by using sound statistical methods. She is vaguely familiar with the classification methods such as logistic regression (LR) and wants to seek your advice.

1. She has a sample of size $n = 199$ and number of variables $d = 27$ and wishes to use LR to build a classifier that corresponds to LR.

   (a) Describe the general form of the classifier that corresponds to LR.

   (b) Write down the form of the likelihood function, and derive a **Newton's method** for solving the maximum likelihood modeling fitting. Clearly state the algorithm resulted from your derivation.

   (c) Describe how you would estimate the classification error of your approach.

   (d) Clearly state the assumptions that are made when LR is used.

2. Realizing that the sample size is rather small relative to the number of variables. We can use ℓ_2 **regularization** to improve the model. How would this change the form of your Newton's method based on algorithm for maximum likelihood modeling fitting?

3. In a different research project, her objective is to study the ecosystem services (e.g. carbon storage, erosion protection) provided by the forest in that region. Hence, in this setting, y_i are *categorical variables* indicating the type of services. Extend your algorithm for this setting (i.e., **multinomial logistic regression**) without considering regularization. Describe the general form of the classifier. Write down the form of the likelihood function, and clearly state the algorithm for modeling fitting.

4. Now apply the algorithm you derived in Part (3) to analyze some real-data sets.

   The data set contains both the training and testing data from a remote sensing study which mapped different forest types based on their spectral characteristics at visible-to-near infrared wavelengths, using ASTER satellite imagery. The output (forest type map) can be used to identify and/or quantify the ecosystem services (e.g. carbon storage, erosion protection) provided by the forest. There are four classes:

   Class: 's' ('Sugi' forest), 'h' ('Hinoki' forest), 'd' ('Mixed deciduous' forest), 'o' ('Other' non-forest land)

   The variables are

   - b1 - b9: ASTER image bands containing spectral information in the green, red, and near infrared wavelengths for three dates (Sep. 26, 2010; March 19, 2011; May 08, 2011.)
   - pred_minus_obs_S_b1 - pred_minus_obs_S_b9: Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 's' class (b1-b9).
   - pred_minus_obs_H_b1 - pred_minus_obs_H_b9: Predicted spectral values (based on spatial interpolation) minus actual spectral values for the 'h' class (b1-b9).

   Fit the model using training data, and report the model you have found. Then perform classification on the test data. In this data set, the classes of the test data are also provided; use them to evaluate the classification error of your method.

5. Now state the gradient of your log-likelihood function. Implement the **gradient descent** method using your derived gradient to maximize the likelihood. Fit the model using training data, and report the model you have found. Then perform classification on the test data. Briefly compare the gradient descent method with Newton's method.

6. Now let's study the **back tracking line search**.

   (a) We can use backtracking approach to appropriately choose step-size, and the sufficient decrease condition to terminate the line search procedure. The optimization problem is

   $$\min_{\theta} f(\theta).$$

   Let p_k be the search direction. In most basic form, backtracking proceeds as follows:

   ```
   i. Choose $\bar{\alpha} > 0$, $\rho, c \in (0,1)$; set $\alpha \leftarrow \bar{\alpha}$;
   ii. repeat until $f(\theta_k + \alpha p_k) \leq f(\theta_k) + c\alpha \nabla f_k^\top p_k$;
   iii. do
           $\alpha \leftarrow \rho\alpha$;
   iv. end (repeat).
   v. Terminate with $\alpha_k = \alpha$
   ```

   Set $\bar{\alpha} = 1$ and choose proper values for ρ and c. Implement Newton's method using this strategy for terminating the line search. Fit the model using training data, and report the model you have found.

7. Now let's implement the **stochastic gradient descent**. Briefly state the algorithm. Set batch-size to be 1, 16, and 32 respectively and compare the performance. Fit the model using training data, and report the model you have found. Then perform classification on the test data.

2 Quadratic Programming [10 pts]

Consider $\min_{x\in\mathbb{R}^n} f(x)$, where $f(x) = \frac{1}{2}(x_1^2 + 2(1-\epsilon)x_1x_2 + x_2^2)$. Find the condition number of the Hessian of f. What happens to the condition number as $\epsilon \to 0$?

3 General Linear models for Softmax regression [20 pts]

Consider a k-classes classification problem, where the probability of belonging to class i is parametrized as follows:

$$p(y = i; \phi) = \phi_i$$

where $\phi_1, \ldots, \phi_k > 0$ specify the probability of each of the outcomes and thus satisfy $\sum_{i=1}^k \phi_i = 1$.

1. Show that this distribution belongs to the exponential family, i.e., show that this probabilistic model can be written in the form

   $$p(y; \phi) = b(y)\exp(\eta^T T(y) - a(\eta))$$

   for some particular function b, T, a.

   Hint: Use $T : \{1, \ldots, k\} \to \mathbb{R}^{k-1}$ given by $T(y)_i = 1\{y = i\}$ where $1\{\cdot\}$ is the indicator function, i.e., it is 1 if its argument is *True* and 0 if its argument is *False.*

2. Compute the response function of the model, i.e., express each ϕ_i in terms of $\{\eta_i\}_{i=1,\ldots,k-1}$.

Suppose we have a k-class classification problem with feature vectors $x^{(i)} \in \mathbb{R}^d$ and associated label $y_i \in \{1, 2, \ldots, k\}$, for all $i = 1, 2, \ldots, n$. We use as probabilistic model the previous softmax model, where parameters η_i are linearly related to the x's, i.e.,

$$\begin{aligned} p(y = i|x; \theta) &= \phi_i \\ &= \frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}} \\ &= \frac{e^{\theta_i^t x}}{\sum_{j=1}^k e^{\theta_j^T x}}. \end{aligned}$$

3. Based on i.i.d observations $x^{(i)}, y_i$, $i = 1, \ldots, n$, compute the log-likelihood function of θ.

4. Compute the gradient of the log-likelihood.