CS 229, Summer 2023
Problem Set #1

Alessandro (`alebarro`)

Due Friday, July 14 at 11:59 pm on Gradescope.

Notes: (1) These questions require thought, but do not require long answers. Please be as concise as possible.

(2) If you have a question about this homework, we encourage you to post your question on our Ed forum, at `https://edstem.org/us/courses/41182/discussion/`.

(3) If you missed the first lecture or are unfamiliar with the collaboration or honor code policy, please read the policy on the course website before starting work.

(4) For the coding problems, you may not use any libraries except those defined in the provided `environment.yml` file. In particular, ML-specific libraries such as scikit-learn are not permitted.

(5) The due date is Friday, July 14 at 11:59 pm. If you submit after Friday, July 14 at 11:59 pm, you will begin consuming your late days. The late day policy can be found in the course website: Course Logistics and FAQ.

All students must submit an electronic PDF version of the written question including plots generated from the codes. We highly recommend typesetting your solutions via LaTeX. All students must also submit a zip file of their source code to Gradescope, which should be created using the `make_zip.py` script. You should make sure to (1) restrict yourself to only using libraries included in the `environment.yml` file, and (2) make sure your code runs without errors. Your submission may be evaluated by the auto-grader using a private test set, or used for verifying the outputs reported in the writeup. Please make sure that your PDF file and zip file are submitted to the corresponding Gradescope assignments respectively. We reserve the right to not give any points to the written solutions if the associated code is not submitted.

Honor code: We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down the solution independently, and without referring to written notes from the joint session. Each student must understand the solution well enough in order to reconstruct it by him/herself. It is an honor code violation to copy, refer to, or look at written or code solutions from a previous year, including but not limited to: official solutions from a previous year, solutions posted online, and solutions you or someone else may have written up in a previous year. Furthermore, it is an honor code violation to post your assignment solutions online, such as on a public git repo. We run plagiarism-detection software on your code against past solutions as well as student submissions from previous years. Please take the time to familiarize yourself with the Stanford Honor Code[1] and the Stanford Honor Code[2] as it pertains to CS courses.

[1] https://communitystandards.stanford.edu/policies-and-guidance/honor-code
[2] https://web.stanford.edu/class/archive/cs/cs106b/cs106b.1164/handouts/honor-code.pdf

1. **[40 points] Linear Classifiers (logistic regression and GDA) (Part a and b can be completed after lecture 2, the rest after lecture 4)**

In this problem, we cover two probabilistic linear classifiers we have covered in class so far. First, a discriminative linear classifier: logistic regression. Second, a generative linear classifier: Gaussian discriminant analysis (GDA). Both of the algorithms find a linear decision boundary that separates the data into two classes, but make different assumptions. Our goal in this problem is to get a deeper understanding of the similarities and differences (and, strengths and weaknesses) of these two algorithms.

For this problem, we will consider two datasets, along with starter codes provided in the following files:

- `src/linearclass/ds1_{train,valid}.csv`
- `src/linearclass/ds2_{train,valid}.csv`
- `src/linearclass/logreg.py`
- `src/linearclass/gda.py`

Each file contains n examples, one example $(x^{(i)}, y^{(i)})$ per row. In particular, the i-th row contains columns $x_1^{(i)} \in \mathbb{R}$, $x_2^{(i)} \in \mathbb{R}$, and $y^{(i)} \in \{0, 1\}$. In the subproblems that follow, we will investigate using logistic regression and Gaussian discriminant analysis (GDA) to perform binary classification on these two datasets.

Typically, a trained model is evaluated by its performance on the validation dataset. The validation dataset is a set of examples drawn from the same (or a similar) distribution as the training data. Intuitively, this is because we need the trained model to correctly predict the label for not only the training data, but also new samples from the same distribution.

(a) [10 points]

In lecture we saw the average empirical loss for logistic regression:

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right),$$

where $y^{(i)} \in \{0, 1\}$, $h_\theta(x) = g(\theta^T x)$ and $g(z) = 1/(1 + e^{-z})$.

Find the Hessian H of this function, and show that for any vector z, it holds true that

$$z^T H z \geq 0.$$

Hint: You may want to start by showing that $\sum_i \sum_j z_i x_i x_j z_j = (x^T z)^2 \geq 0$. Recall also that $g'(z) = g(z)(1 - g(z))$.

Remark: This is one of the standard ways of showing that the matrix H is positive semi-definite, written "$H \succeq 0$." This implies that J is convex, and has no local minima other than the global one. If you have some other way of showing $H \succeq 0$, you're also welcome to use your method instead of the one above.

Answer: The gradient (first derivative) $\nabla_\theta J(\theta)$ of the starting logistic regression's loss function $J(\theta)$ looks like this

$$\nabla_\theta J(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta [y^{(i)} \log(h_\theta(x^{(i)}))] + \nabla_\theta [(1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})] \tag{1}$$

$$\nabla_\theta J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\Big(y^{(i)} - h_\theta(x)\Big)x^{(i)} \tag{2}$$

For simplicity, we will refer to $h_\theta(x)$ as the CRF $g(z)$, where g the sigmoid function, $z = \theta^T x$. Let's take the first derivative of $g(z)$ that will help us in the calculation of the hessian

$$\frac{d}{dz}g(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) = g(z)(1-g(z)) \tag{3}$$

Now let's finally find the hessian H, the matrix of second derivatives

$$H = \nabla_\theta^2 J(\theta) = \nabla_\theta \frac{1}{n}\sum_{i=1}^{n}\Big(y^{(i)} - h_\theta(x)\Big)x^{(i)} \tag{4}$$

$$= \sum_{i=1}^{n} x^{(i)}x^{(i)}g(z)(1-g(z)) \tag{5}$$

H corresponds to the following expression in vector notation

$$H = \nabla_\theta^2 J(\theta) = X^T Z X \tag{6}$$

Where $X \in \mathbb{R}^{d\times n}$ is the matrix of the indipendent variables and $Z \in \mathbb{R}^{d\times d}$ diagonal matrix s.t. $Z_{ii} = g(z^{(i)})(1-g(z^{(i)}))$ are the components of the diagonal.
Suppose we have the following quadratic form

$$q(z) = z^T H z \tag{7}$$

In order to show that $q(z) \geq 0$, we can consider its vectorial subspace $\ker H = \{z \in \mathbb{R}^d | q(z) = \vec{0}\}$ and realize that, since $Z_{ii} \geq 0\ \forall_i$ and the outer product $X^T X \geq 0$ by definition, we can easily observe that $Hz = 0 \iff z = \vec{0}$, hence the $z^T H z \geq 0$ (PSD)

(b) [5 points] **Coding problem.** Follow the instructions in `src/linearclass/logreg.py` to train a logistic regression classifier using Newton's Method. Starting with $\theta = \vec{0}$, run Newton's Method until the updates to θ are small: Specifically, train until the first iteration k such that $\|\theta_k - \theta_{k-1}\|_1 < \epsilon$, where $\epsilon = 1\times 10^{-5}$. Make sure to write your model's predicted probabilities on the validation set to the file specified in the code.

Include a plot of the **validation data** with x_1 on the horizontal axis and x_2 on the vertical axis. To visualize the two classes, use a different symbol for examples $x^{(i)}$ with $y^{(i)} = 0$ than for those with $y^{(i)} = 1$. On the same figure, plot the decision boundary found by logistic regression (i.e., line corresponding to $p(y|x) = 0.5$).

Note: If you want to print the loss during training, you may encounter some numerical instability issues. Recall that the loss function on an example (x, y) is defined as

$$y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x)),$$

where $h_\theta(x) = (1+\exp(-x^\top\theta))^{-1}$. Technically speaking, $h_\theta(x) \in (0,1)$ for any $\theta, x \in \mathbb{R}^d$. However, in Python a real number only has finite precision. So it is possible that in your implementation, $h_\theta(x) = 0$ or $h_\theta(x) = 1$, which makes the loss function ill-defined. A typical solution to the numerical instability issue is to add a small perturbation. In this case, you can compute the loss function using

$$y\log(h_\theta(x)+\epsilon) + (1-y)\log(1-h_\theta(x)+\epsilon),$$

instead, where ϵ is a very small perturbation (for example, $\epsilon = 10^{-5}$). **Answer:**

(c) [5 points] Recall that in GDA we model the joint distribution of (x, y) by the following equations:

$$p(y) = \begin{cases} \phi & \text{if } y = 1 \\ 1 - \phi & \text{if } y = 0 \end{cases} \tag{8}$$

$$p(x|y=0) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \tag{9}$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right),$$

where ϕ, μ_0, μ_1, and Σ are the parameters of our model.

Suppose we have already fit ϕ, μ_0, μ_1, and Σ, and now want to predict y given a new point x. To show that GDA results in a classifier that has a linear decision boundary, show the posterior distribution can be written as

$$p(y = 1 \mid x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-(\theta^T x + \theta_0))},$$

where $\theta \in \mathbb{R}^d$ and $\theta_0 \in \mathbb{R}$ are appropriate functions of ϕ, Σ, μ_0, and μ_1. State the value of θ and θ_0 as a function of $\phi, \mu_0, \mu_1, \Sigma$ explicitly.

Answer: Let's first apply the Bayes rule and we obtain

$$P(y=1|x) = \frac{P(y=1)P(x|y=1)}{P(X)} \tag{10}$$

Where $P(x) = P(y=0)P(x|y=0) + P(y=1)P(x|y=1)$. By substituting the known expression from hypothesis, and imposing $K = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}$, we obtain

$$P(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{\phi K \exp\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\}}{(1-\phi)K\exp\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\} + \phi K\exp\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\}} \tag{11}$$

Now, let's define two temporary constants α and β

$$\alpha = -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \log\phi$$

$$\beta = -\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - \log(1-\phi)$$

And by plugging those into out equation we obtain

$$P(y=1|x;\phi,\mu_0,\mu_1,\Sigma) = \frac{1}{1+exp[\alpha-\beta]} \tag{12}$$

Let's better analyze $\alpha - \beta$

$$\alpha - \beta = -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \log\phi + \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - \log(1-\phi) \tag{13}$$

And by defining the GDA's model parameters as functions of $\phi, \mu_0, \mu_1, \Sigma$

$$\theta = \Sigma^{-1}(\mu_1 - \mu_0) \tag{14}$$

$$\theta_0 = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log\frac{(1-\phi)}{\phi} \tag{15}$$

We have successfully proven that

$$P(y = 1|x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-\theta^T x - \theta_0)} \tag{16}$$

GDA results in a classifier with linear decision boundary and matches the form of logistic regression even though parameters are differently defined.

(d) [7 points] Given the dataset, we claim that the maximum likelihood estimates of the parameters are given by

$$\phi = \frac{1}{n} \sum_{i=1}^{n} 1\{y^{(i)} = 1\} \tag{17}$$

$$\mu_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}} \tag{18}$$

$$\mu_1 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}} \tag{19}$$

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

The log-likelihood of the data is

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi). \end{aligned} \tag{20}$$

By maximizing ℓ with respect to the four parameters, prove that the maximum likelihood estimates of ϕ, μ_0, μ_1, and Σ 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.)

Answer: Let $l(\theta) = \log L(\theta)$ and let's define the log likelihood function as follows

$$l(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} P(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) P(y^{(i)}; \phi) \tag{21}$$

$$= \sum_{i=1}^{n} \log P(x^{(i)|y^{(i)}; \mu_0, \mu_1, \Sigma} + \log P(y^{(i)}; \phi) \tag{22}$$

Assuming that there is at least one positive and one negative example, we will proceed step by step, taking the gradient $\nabla_k l(\phi, \mu_0, \mu_1, \Sigma)$ w.r.t. each variable and setting it to zero to obtain the closed-form expressions.

1. Σ

$$\nabla_\Sigma \log P(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) = 0 \tag{23}$$

$$\nabla_\Sigma \left[\log 1 - \log(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}} \right] - \nabla_\Sigma \left[\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] = 0 \tag{24}$$

$$-\frac{1}{2}\nabla_\Sigma|\Sigma| - \frac{1}{2}\nabla_\Sigma(x-\mu)^T\Sigma^{-1}(x-\mu) = 0 \tag{25}$$

$$\Sigma = (x-\mu)(x-\mu)^T = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \tag{26}$$

2. $\mu_{0,1}$

$$-\nabla_{\mu_{0,1}}\frac{1}{2}(x-\mu_{0,1})^T\Sigma^{-1}(x-\mu_{0,1}) = 0 \tag{27}$$

$$= \Sigma^{-1}(\mu_{0,1} - x) = \frac{\sum_i I\{y^{(i)} = 0,1\}x^{(i)}}{\sum_i I\{y^{(i)} = 0,1\}} = \mu_{0,1} \tag{28}$$

3. ϕ

$$\nabla_\phi y\log\phi + (1-y)\nabla_\phi\log(1-\phi) = 0 \tag{29}$$

$$= \sum_i \frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi} = 0 \tag{30}$$

$$\frac{1}{n}\sum_i I\{y^{(i)=1}\} = \phi \tag{31}$$

(e) [5 points] **Coding problem.** In `src/linearclass/gda.py`, fill in the code to calculate ϕ, μ_0, μ_1, and Σ, use these parameters to derive θ, and use the resulting GDA model to make predictions on the validation set. Make sure to write your model's predictions on the validation set to the file specified in the code.

Include a plot of the **validation data** with x_1 on the horizontal axis and x_2 on the vertical axis. To visualize the two classes, use a different symbol for examples $x^{(i)}$ with $y^{(i)} = 0$ than for those with $y^{(i)} = 1$. On the same figure, plot the decision boundary found by GDA (i.e, line corresponding to $p(y|x) = 0.5$).

Answer:

(f) [2 points] For Dataset 1, compare the validation set plots obtained in part (b) and part (e) from logistic regression and GDA respectively, and briefly comment on your observation in a couple of lines.

Answer: For Dataset 1, as for logistic regression, the data seems to be plotted in a scattered way, and the classifier does an overall good job. As for GDA, data is plotted and concentrated in a horizontal line, not a very optimal scenario for a linear classifier that might be a bit underperforming.

(g) [5 points] Repeat the steps in part (b) and part (e) for Dataset 2. Create similar plots on the **validation set** of Dataset 2 and include those plots in your writeup.

On which dataset does GDA seem to perform worse than logistic regression? Why might this be the case?

Answer: As previously mentioned, GDA seems to be underperforming in Dataset 1. In fact, as for Dataset 2, the two models seems to have a very similar performance in terms of classification. This could be because, GDA needs some requireres some particular data policies, such as the two example groups needs to be normally distributed by themselves. If we remove some data points from the Dataset 2 though, the GDA will definitely shine and perform much better than Logistic regression.

(h) **[1 points]** For the dataset where GDA performed worse in parts (f) and (g), can you find a transformation of the $x^{(i)}$'s such that GDA performs significantly better? What might this transformation be?

Answer: A simple quadratic or polynomial transformation could significantly improve the performance of the GDA model in Dataset 1.

2. **[25 points] Poisson Regression (can be completed after lecture 3)**

In this question we will construct another kind of a commonly used GLM, which is called Poisson Regression. In a GLM, the choice of the exponential family distribution is based on the kind of problem at hand. If we are solving a classification problem, then we use an exponential family distribution with support over discrete classes (such as Bernoulli, or Categorical). Simiarly, if the output is real valued, we can use Gaussian or Laplace (both are in the exponential family). Sometimes the desired output is to predict counts, for example, predicting the number of emails expected in a day, or the number of customers expected to enter a store in the next hour, etc. based on input features (also called covariates). You may recall that a probability distribution with support over integers (i.e., counts) is the Poisson distribution, and it also happens to be in the exponential family.

In the following sub-problems, we will start by showing that the Poisson distribution is in the exponential family, derive the functional form of the hypothesis, derive the update rules for training models, and finally using the provided dataset to train a real model and make predictions on the test set.

(a) [5 points] Consider the Poisson distribution parameterized by λ:

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

(Here y has positive integer values and $y!$ is the factorial of y.) Show that the Poisson distribution is in the exponential family, and clearly state the values for $b(y)$, η, $T(y)$, and $a(\eta)$.

Answer: In order to show that the Poisson probability distribution is part of the exponential family, we need to rewrite the expression and highlight the searched parameters

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{e^{-\lambda}e^{y\log\lambda}}{y!} = \frac{1}{y!}e^{y\log\lambda - e^{\log\lambda}} \tag{32}$$

Now we can clearly see the values for $b(y)$, η, $T(y)$ and $a(\eta)$
$\eta = \log\lambda$
$b(y) = \frac{1}{y!}$
$T(y) = y$
$a(\eta) = e^{\log\lambda}$

(b) [3 points] Consider performing regression using a GLM model with a Poisson response variable. What is the canonical response function for the family? (You may use the fact that a Poisson random variable with parameter λ has mean λ.)

Answer: From the previous exercised we derived canonical and natural parameters of the Poisson's regression related to the GLM expression (exponential family). The canonical response function g is some function such that

$$h_\theta(x) = g(\eta) \tag{33}$$

We found out that η and the Poisson's random variable are linked by the following relationship

$$\eta = \log\lambda \tag{34}$$

So, solving the equation by the random variable λ (and no less than the mean of the probability distribution), we obtain

$$\lambda = e^\eta = g(\eta) \tag{35}$$

which is the canonical response function.

(c) [7 points] For a training set $\{(x^{(i)}, y^{(i)}); i = 1, \ldots, n\}$, let the log-likelihood of an example be $\log p(y^{(i)}|x^{(i)}; \theta)$. By taking the derivative of the log-likelihood with respect to θ_j, derive the stochastic gradient ascent update rule for learning using a GLM model with Poisson responses y and the canonical response function.

Answer: In order to ultimately define the stochastic gradient ascent update rule, we need to take in consideration the Poisson's probability density function and the corresponding canonical response function

$$\log L(\theta) = l(\theta) = \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y}}{y!} \tag{36}$$

Let's substitute the random variable λ with the related CRF value

$$= \sum_{i=1}^{n} \frac{\log e^{e^{\theta^T x^{(i)}}} (e^{\theta^T x^{(i)}})^{y}}{(y^{(i)})!} = \sum_{i=1}^{n} -e^{\theta^T x^{(i)}} + y^{(i)}(\theta^T x^{(i)}) - \log(y^{(i)})! \tag{37}$$

Now taking the gradient of the log likelihood function, and specifically differentiating with respect to $\theta^{(j)}$ we obtain

$$\frac{\partial}{\partial \theta^{(j)}} \left[\sum_{i=1}^{n} -e^{\theta^T x^{(i)}} + y^{(i)}(\theta^T x^{(i)}) - \log(y^{(i)})! \right] \tag{38}$$

$$\sum_{i=1}^{n} x^{(j)}[y^(j) - e^{\theta^T x^{(j)}}] \tag{39}$$

$$\nabla_\theta l(\theta) = X^T[y - e^{X\theta}] \tag{40}$$

But since we are considering stochastic gradient ascent, we shouldn't consider the whole gradient, but only one random example (r) over the entire example range and iterate it until convergence. The final SGA update rule, will look like this

$$\theta^{(k)} = \theta^{(k-1)} + \alpha[x^{(r)}(y^{(r)} - e^{\theta^T x^{(r)}})] \tag{41}$$

With $r =$ "random index included from the total number of examples"

(d) [10 points] **Coding problem**

Consider a website that wants to predict its daily traffic. The website owners have collected a dataset of past traffic to their website, along with some features which they think are useful in predicting the number of visitors per day. The dataset is split into train/valid sets and the starter code is provided in the following files:

- `src/poisson/{train,valid}.csv`
- `src/poisson/poisson.py`

We will apply Poisson regression to model the number of visitors per day. Note that applying Poisson regression in particular assumes that the data follows a Poisson distribution whose natural parameter is a linear combination of the input features (*i.e.,* $\eta = \theta^T x$). In `src/poisson/poisson.py`, implement Poisson regression for this dataset and use *full batch gradient ascent* to maximize the log-likelihood of θ. For the stopping criterion, check if the change in parameters has a norm smaller than a small value such as 10^{-5}.

Using the trained model, predict the expected counts for the **validation set**, and create a scatter plot between the true counts vs predicted counts (on the validation set). In the

scatter plot, let x-axis be the true count and y-axis be the corresponding predicted expected count. Note that the true counts are integers while the expected counts are generally real values.

Answer:

3. **[15 points] Convexity of Generalized Linear Models (can be completed after lecture 3)**

In this question we will explore and show some nice properties of Generalized Linear Models, specifically those related to its use of Exponential Family distributions to model the output.

Most commonly, GLMs are trained by using the negative log-likelihood (NLL) as the loss function. This is mathematically equivalent to Maximum Likelihood Estimation (*i.e.,* maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood). In this problem, our goal is to show that the NLL loss of a GLM is a *convex* function w.r.t the model parameters. As a reminder, this is convenient because a convex function is one for which any local minimum is also a global minimum, and there is extensive research on how to optimize various types of convex functions efficiently with various algorithms such as gradient descent or stochastic gradient descent.

To recap, an exponential family distribution is one whose probability density can be represented as

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

where η is the *natural parameter* of the distribution. Moreover, in a Generalized Linear Model, η is modeled as $\theta^T x$, where $x \in \mathbb{R}^d$ are the input features of the example, and $\theta \in \mathbb{R}^d$ are learnable parameters. In order to show that the NLL loss is convex for GLMs, we break down the process into sub-parts, and approach them one at a time. Our approach is to show that the second derivative (*i.e.,* Hessian) of the loss w.r.t the model parameters is Positive Semi-Definite (PSD) at all values of the model parameters. We will also show some nice properties of Exponential Family distributions as intermediate steps.

For the sake of convenience we restrict ourselves to the case where η is a scalar. Assume $p(Y|X;\theta) \sim \text{ExponentialFamily}(\eta)$, where $\eta \in \mathbb{R}$ is a scalar, and $T(y) = y$. This makes the exponential family representation take the form

$$p(y;\eta) = b(y)\exp(\eta y - a(\eta)).$$

Note that the above probability density is for a single example (x, y).

(a) [5 points] Derive an expression for the mean of the distribution. Show that $\mathbb{E}[Y;\eta] = \frac{\partial}{\partial \eta} a(\eta)$ (note that $\mathbb{E}[Y;\eta] = \mathbb{E}[Y|X;\theta]$ since $\eta = \theta^T x$). In other words, show that the mean of an exponential family distribution is the first derivative of the log-partition function with respect to the natural parameter.

Hint: Start with observing that $\frac{\partial}{\partial \eta}\int p(y;\eta)dy = \int \frac{\partial}{\partial \eta} p(y;\eta)dy$.

Answer: Let's start by observing that

$$\frac{\partial}{\partial \eta}\int P(y;\eta)dy = \int \frac{\partial}{\partial \eta} P(y;\eta)dy \tag{42}$$

The derivative of $P(y;\eta)$ w.r.t. η

$$\frac{\partial}{\partial \eta} P(y;\eta) = b(y)\exp[n\eta - a(\eta)]\left(y - \frac{\partial}{\partial \eta}a(\eta)\right) = P(y;\eta)\left(y - \frac{\partial}{\partial \eta}\right) \tag{43}$$

Going back to the integral, and knowing that $\int P(y;\eta)dy = 1$

$$\int P(y;\eta)\left(y - \frac{\partial}{\partial \eta}a(\eta)\right)dy = \frac{\partial}{\partial \eta}\int P(y;\eta)dy = 0 \tag{44}$$

$$\int P(y;\eta)ydy = \frac{\partial}{\partial\eta}a(\eta)\int P(y;\eta)dy \tag{45}$$

Coming to the following conclusion

$$\mathbb{E}[Y;\eta] = \frac{\partial}{\partial\eta}a(\eta) \tag{46}$$

(b) [5 points] Next, derive an expression for the variance of the distribution. In particular, show that $\text{Var}(Y;\eta) = \frac{\partial^2}{\partial\eta^2}a(\eta)$ (again, note that $\text{Var}(Y;\eta) = \text{Var}(Y|X;\theta)$). In other words, show that the variance of an exponential family distribution is the second derivative of the log-partition function w.r.t. the natural parameter.

Hint: Building upon the result in the previous sub-problem can simplify the derivation.

Answer: In order to show that $\text{Var}(Y;\eta) = \frac{\partial^2}{\partial^2\eta}a(\eta)$, let's first recall the definition of variance for probability density functions

$$\text{Var}(Y;\eta) = \mathbb{E}[Y^2;\eta] - \mathbb{E}[Y;\eta]^2 \tag{47}$$

Now, with the results obtained from the previous exercise, we can impose

$$\frac{\partial}{\partial\eta}\mathbb{E}[Y;\eta] = \frac{\partial^2}{\partial^2\eta}a(\eta) \tag{48}$$

And starting from here, let's explicitly demonstrate that $\frac{\partial}{\partial\eta}\mathbb{E}[Y;\eta]$ corresponds to $\text{Var}(Y;\eta)$

$$\frac{\partial}{\partial\eta}\int P(y;\eta)ydy = \int yP(y;\eta)\left(y - \frac{\partial}{\partial\eta}a(\eta)\right)dy \tag{49}$$

$$= \int y^2P(y;\eta)dy - \int yP(y;\eta)\frac{\partial}{\partial\eta}a(\eta)dy \tag{50}$$

$$= \mathbb{E}[Y^2;\eta] - \mathbb{E}[Y;\eta]^2 = \text{Var}(Y;\eta) \tag{51}$$

Hence, we can conclude that

$$\text{Var}(Y;\eta) = \mathbb{E}[Y^2;\eta] - \mathbb{E}[Y;\eta]^2 = \frac{\partial^2}{\partial^2\eta}a(\eta) \tag{52}$$

(c) [5 points] Finally, write out the loss function $\ell(\theta)$, the NLL of the distribution, as a function of θ. Then, calculate the Hessian of the loss w.r.t θ, and show that it is always PSD. This concludes the proof that NLL loss of GLM is convex.

Hint 1: Use the chain rule of calculus along with the results of the previous parts to simplify your derivations.

Hint 2: Recall that variance of any probability distribution is non-negative.

Answer: We can define the NLL as follows

$$J(\eta) = -\log\prod_{i=1}^{n}P(y;\eta) = -\sum_{i=1}^{n}\log[b(y^{(i)})] + \eta y - a(\eta) \tag{53}$$

From hypothesis, we know that $\eta = \theta^T x$

$$J(\theta) = -\log \prod_{i=1}^{n} P(y;\theta) = -\sum_{i=1}^{n} \log[b(y^{(i)})] + (\theta^T x^{(i)}) y^{(i)} - a(\theta^T x^{(i)}) \tag{54}$$

Let's first compute the gradient $\nabla_\theta J(\theta)$

$$\nabla_\theta J(\theta) = -\sum_{i=1}^{n} x^{(i)} y^{(i)} - x^{(i)} \frac{\partial}{\partial \theta} a(\theta^T x^{(i)}) \tag{55}$$

And we can re-write this as

$$\nabla_\theta J(\theta) = \sum_{i=1}^{n} x^{(i)} \left[\mathbb{E}[Y; \theta^T x^{(i)}] - y^{(i)} \right] \tag{56}$$

Now we can calculate the hessian H as

$$H = \nabla_\theta^2 J(\theta) \tag{57}$$

By computing and substituting known relationships obtained from the previous points

$$\nabla_\theta \sum_{i=1}^{n} x^{(i)} \left[\frac{\partial}{\partial \theta} a(\theta^T x^{(i)}) - y^{(i)} \right] = \sum_{i=0}^{n} x^{(i)} x^{(i)} \frac{\partial^2}{\partial^2 \theta} a(\theta^T x^{(i)}) \tag{58}$$

$$H = X X^T \mathsf{Var}[Y; \theta^T x] \tag{59}$$

Since the outer-product $XX^T \geq 0$ and $\mathsf{Var}[Y; \theta^T x] \geq 0$, it is trivial to show that

$$H \geq 0 \quad (PSD) \tag{60}$$

Therefore, in conclusion, the NLL loss of GLMs is convex.

Remark: The main takeaways from this problem are:

- Any GLM model is convex in its model parameters.
- The exponential family of probability distributions are mathematically nice. Whereas calculating mean and variance of distributions in general involves integrals (hard), surprisingly we can calculate them using derivatives (easy) for exponential family.

4. **[25 points] Locally weighted linear regression (can be completed after lecture 2)**

(a) [10 points] Consider a linear regression problem in which we want to "weight" different training examples differently. Specifically, suppose we want to minimize

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n} w^{(i)} \left(\theta^T x^{(i)} - y^{(i)}\right)^2.$$

In class, we worked out what happens for the case where all the weights (the $w^{(i)}$'s) are the same. In this problem, we will generalize some of those ideas to the weighted setting. We will assume $w^{(i)} > 0$ for all i.

i. [2 points] Show that $J(\theta)$ can also be written

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

for an appropriate matrix W, and where X and y are as defined in class. Clearly specify the value of each element of the matrix W.

ii. [4 points] If all the $w^{(i)}$'s equal 1, then we saw in class that the normal equation is

$$X^T X\theta = X^T y,$$

and that the value of θ that minimizes $J(\theta)$ is given by $(X^T X)^{-1} X^T y$. By finding the derivative $\nabla_\theta J(\theta)$ and setting that to zero, generalize the normal equation to this weighted setting, and give the new value of θ that minimizes $J(\theta)$ in closed form as a function of X, W and y.

iii. [4 points] Suppose we have a dataset $\{(x^{(i)}, y^{(i)}); i = 1\ldots, n\}$ of n independent examples, but we model the $y^{(i)}$'s as drawn from conditional distributions with different levels of variance $(\sigma^{(i)})^2$. Specifically, assume the model

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

That is, each $y^{(i)}$ is drawn from a Gaussian distribution with mean $\theta^T x^{(i)}$ and variance $(\sigma^{(i)})^2$ (where the $\sigma^{(i)}$'s are fixed, known, constants). Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem. State clearly what the $w^{(i)}$'s are in terms of the $\sigma^{(i)}$'s.

In other words, this suggests that if we have prior knowledge on the noise levels (the variance of the label $y^{(i)}$ conditioned on $x^{(i)}$) of all the examples, then we should use weighted least squares with weights depending on the variances.

Answer:

i. Let's first well-define the elements that compose our expression. $X \in \mathbb{R}^{d\times d}$ represents all the examples in out data set, $y \in \mathbb{R}_d$ contains the dependent variables.
$w^{(i)} \in \mathbb{R}^{d+1}, \mathbf{w}^{(i)} = (0, \ldots, w_{ij}, \ldots, 0)$ is the local weight vector that re-estimates the output from the squared loss function for every example. All the vectors can be summarized in $W \in \mathbb{R}^{d+1\times d+1}$ and looks like this

$$W = \begin{bmatrix} w^{(11)} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & w^{(d+1\ d+1)} \end{bmatrix}$$

Hence W is a diagonal matrix and $W \geq 0$ (PSD).
From this assumptions we can re-write $J(\theta)$ in vector notation from sum notation.

$$J(\theta) = \frac{1}{2}\{w^{(1)}(\theta^T x^{(1)} - y^{(1)})^2 + \cdots + w^{(n)}(\theta^T x^{(n)} - y^{(n)})^2\} \tag{61}$$

Let $z \in \mathbb{R}^n$ be a vector and $z^{(i)} = (\theta^T x^{(i)} - y^{(i)})$

$$J(\theta) = \frac{1}{2} z^T \sum_{i=1}^{n} w^{(i)} z^{(i)} = z^T W z \tag{62}$$

And finally obtaining the final quadratic form

$$J(\theta) = (X\theta - y)^T W (X\theta - y) \tag{63}$$

Note that $\frac{1}{2}$ is a mathematical convention to simplify the differentiation and does not affect the final result (can be removed).

ii. In order to find the optimal value of $\hat{\theta}$, we need to compute the gradient with respect to θ and set it to zero

$$\nabla_\theta J(\theta) = 0 \tag{64}$$

Let's differentiate the expression with respect to θ_i the i-th vector component.

$$\frac{\partial}{\partial \theta_i} \frac{1}{2} \sum_{i=1}^{n} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2 \tag{65}$$

$$= X^T \left(\sum_{i=1}^{n} w^{(i)} \theta^T x^{(i)} - \sum_{i=1}^{n} w^{(i)} y^{(i)} \right) = X^T (WX\theta - WY) = X^T W X \theta - WYX = 0 \tag{66}$$

Finally coming to the conclusion that the generalized normal equation for a LWR model is

$$\hat{\theta} = WYX(X^T W X)^{-1} \tag{67}$$

iii. In first place, we need to define $L(\theta)$ the likelihood function. For mathematical convenience, we will consider $l(\theta) = \log L(\theta)$.

$$L(\theta) = \prod_{i=1}^{n} P(y^{(i)} | x^{(i)}; \theta), l(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} P(y^{(i)} | x^{(i)}; \theta) \tag{68}$$

$$= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi}\sigma^{(i)}} \right) - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} = nK - \sum_{i=1}^{n} \log \sigma^{(i)} + \frac{[y^{(i)} - \theta^T x^{(i)}]}{2(\sigma^{(i)})^2} \tag{69}$$

Now, let's find the optimal value $\ddot{\theta}$ and highlight the value of $w^{(i)}|_{\sigma^{(i)}}$ by taking the gradient, maximizing the likelihood $l(\theta)$ and setting it to 0

$$\nabla_\theta l(\theta) = 0 \tag{70}$$

$$\frac{\partial}{\partial\theta^{(i)}}\left[nK-\left(\sum_{i=1}^{n}\log\sigma^{(i)}+\sum_{i=1}^{n}\frac{[y^{(i)}-\theta^{T}x^{(i)}]^{2}}{2(\sigma^{(i)})^{2}}\right)\right]=\sum_{i=1}^{n}\frac{[y^{(i)}-\theta^{T}x^{(i)}]x^{(i)}}{(\sigma^{(i)})^{2}}=0 \tag{71}$$

If we assume $w^{(i)}=\frac{1}{(\sigma^{(i)})^{2}}$ than, all we are left is the normal equation for LWR models and the related matrix $W=(w^{(i)}\dots w^{(n)})$

$$X^{T}[W(Y-X\theta)]=0 \tag{72}$$

And starting from the idea to maximize the likelihood we ended up on a weighted linear regression problem

$$\ddot{\theta}=WX^{T}Y(X^{T}WX)^{-1} \tag{73}$$

(b) [10 points] **Coding problem.**

We will now consider the following dataset (the formatting matches that of Datasets 1-4, except $x^{(i)}$ is 1-dimensional):

`src/lwr/{train,valid,test}.csv`

In `src/lwr/lwr.py`, implement locally weighted linear regression using the normal equations you derived in Part (a) and using

$$w^{(i)}=\exp\left(-\frac{\|x^{(i)}-x\|_{2}^{2}}{2\tau^{2}}\right).$$

Train your model on the `train` split using $\tau=0.5$, then run your model on the `valid` split and report the mean squared error (MSE). Finally plot your model's predictions on the validation set (plot the training set with blue 'x' markers and the predictions on the validation set with a red 'o' markers). Does the model seem to be under- or overfitting?
Answer: It seems to underfit the data, at some instances it's too approximative and the value of MSE is high on average (≈ 0.33).

(c) [5 points] **Coding problem.**

We will now tune the hyperparameter τ. In `src/lwr/tau.py`, find the MSE value of your model on the validation set for each of the values of τ specified in the code. For each τ, plot your model's predictions on the validation set in the format described in part (b). Report the value of τ which achieves the lowest MSE on the `valid` split, and finally report the MSE on the `test` split using this τ-value.

Answer: Now the model nicely fits the data and the MSE ≈ 0.0124, with $\tau=0.05$.

5. **[10 points] Linear invariance of optimization algorithms (can be completed after lecture 2)**

Consider using an iterative optimization algorithm (such as Newton's method, or gradient descent) to minimize some continuously differentiable function $f(x)$. Suppose we initialize the algorithm at $x^{(0)} = \vec{0}$. When the algorithm runs, it will produce a value of $x \in \mathbb{R}^d$ for each iteration: $x^{(1)}, x^{(2)}, \ldots$.

Now, let some non-singular square matrix $A \in \mathbb{R}^{d \times d}$ be given, and define a new function $g(z) = f(Az)$. Consider using the same iterative optimization algorithm to optimize g (with initialization $z^{(0)} = \vec{0}$). If the values $z^{(1)}, z^{(2)}, \ldots$ produced by this method necessarily satisfy $z^{(i)} = A^{-1}x^{(i)}$ for all i, we say this optimization algorithm is **invariant to linear reparameterizations**.

(a) [7 points] Show that Newton's method (applied to find the minimum of a function) is invariant to linear reparameterizations. Note that since $z^{(0)} = \vec{0} = A^{-1}x^{(0)}$, it is sufficient to show that if Newton's method applied to $f(x)$ updates $x^{(i)}$ to $x^{(i+1)}$, then Newton's method applied to $g(z)$ will update $z^{(i)} = A^{-1}x^{(i)}$ to $z^{(i+1)} = A^{-1}x^{(i+1)}$.[3]

Answer: Let's first recall the expression for Newton's method with respect to $x^{(i)}$ and a function $f(x^{(i)})$ as update parameters

$$x^{(i+1)} = x^{(i)} - \frac{f(x^{(i)})}{f'(x^{(i)})} \tag{74}$$

Given the initial conditions

$$x^{(0)} = \vec{0}, z^{(0)} = \vec{0} \tag{75}$$

And knowing that

$$g(z) = f(Az), z^{(i)} = A^{-1}x^{(i)} \tag{76}$$

$$z^{(i+1)} = z^{(i)} - \frac{g(z^{(i)})}{g'(z^{(i)})} = z^{(i)} - \frac{f(Az^{(i)})}{Af'(Az^{(i)})}. \tag{77}$$

Substituting $z^{(i)} = A^{-1}x^{(i)}$ gives us

$$A^{-1}x^{(i+1)} = A^{-1}x^{(i)} - \frac{f(x^{(i)})}{f'(x^{(i)})}A^{-1}. \tag{78}$$

By multiplying both sides by A, A^{-1} simplifies. We come to the conclusion that

$$x^{(i+1)} = x^{(i)} - \frac{f(x^{(i)})}{f'(x^{(i)})}, \tag{79}$$

Which finally shows that the relationship $z^{(i)} = A^{-1}x^{(i)}$ is valid $\forall i$: Newton's method is invariant to linear reparameterizations.

(b) [3 points] Is gradient descent invariant to linear reparameterizations? Justify your answer.

Answer: Let's recall gradient descent's expression, and checking if updating the variables $x^{(i)}$ and $z^{(i)}$ respectively returns us the same result

$$k^{(i+1)} = k^{(i)} - \alpha f'(k^{(i+1)}) \tag{80}$$

[3] Note that for this problem, you must explicitly prove any matrix calculus identities that you wish to use that are not given in the lecture notes.

Applying GD's update rule with both $x^{(i)}$ and $z^{(i)}$

$$x^{(i+1)} = x^{(i)} - \alpha f'(x^{(i+1)}) \tag{81}$$

$$z^{(i+1)} = z^{(i)} - \alpha (g(z^{(i+1)}))' \tag{82}$$

Let's now consider the last example and express $z^{(i)}$ and $g(z)$ in terms of $x^{(i)}$ and $f(x)$ by substituting the given relationships

$$A^{-1}x^{(i+1)} = A^{-1}x^{(i)} - \alpha f'(x^{(i)})A \tag{83}$$

Multiplying both sides of the equations by A leads us to

$$x^{(i+1)} = x^{(i)} - \alpha f'(x^{(i)})A^2 \tag{84}$$

Coming to the final conclusion

$$f'(x^{(i)})A^2 = f'(x^{(i)}) \tag{85}$$

Which restrincts the validity of our update rule to a case such that $d = 1$. Therefore gradient descent is subsceptible to linear reparamaterization $\forall d > 1$.