CS 229, Spring 2023
Problem Set #1

Anthony Weng (ad2weng)

Due Wednesday, April 26 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/37893/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 Wednesday, April 26 at 11:59 pm. If you submit after Wednesday, April 26 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. **[25 points] Poisson Regression**

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.

1(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: First observe the following:

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

$$= \frac{1}{y!} \times \exp(\log \lambda y - \lambda) \tag{2}$$

Equation 2 allows us to more clearly identify the constituent components of an exponential family distribution's PDF; i.e., $p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$:

$$b(y) = \frac{1}{y!} \tag{3}$$

$$\eta = \log \lambda \tag{4}$$

$$T(y) = y \tag{5}$$

$$a(\eta) = \lambda = e^{\eta} \tag{6}$$

where the right-hand most expression of Equation 6 confirms that $a(\eta)$ is indeed a function of η. Thus, we have demonstrated that the Poisson distribution is in the exponential family ■.

1(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: The canonical response function is the function g which provides the distribution's mean (here, λ) as a function of η. We observe the following:

$$g(\eta) = \mathbb{E}[T(y); \eta] \tag{7}$$
$$= \lambda \tag{8}$$
$$= e^{\eta} \tag{9}$$

where the last line follows from $e^{(\eta = \log \lambda)} = \lambda = \mathbb{E}[T(y); \eta]$. Thus, the canonical response function for a Poisson response variable is $g(\eta) = e^{\eta}$ ■.

1(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: First, let us encode the log-likelihood as a function of θ:

$$\ell(\theta) = \log p(y^{(i)} \mid x^{(i)}; \theta) \tag{10}$$

$$= \log \left(\frac{1}{y^{(i)}!} \times \exp(\log \lambda y^{(i)} - \lambda) \right) \tag{11}$$

$$= \log \frac{1}{y^{(i)}!} \times \exp(\eta y^{(i)} - e^{\eta}) \tag{12}$$

$$= \log \frac{1}{y^{(i)}!} \times \exp(\theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}}) \tag{13}$$

$$= \log \frac{1}{y^{(i)}} + \theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} \tag{14}$$

Next, we compute the derivative of $\ell(\theta)$ w.r.t θ_j:

$$\nabla \ell(\theta)_{\theta_j} = \frac{\partial}{\partial_{\theta_j}} \left(\log \frac{1}{y^{(i)}!} + \theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} \right) \tag{15}$$

$$= 0 + x_j^{(i)} y^{(i)} - e^{\theta^T x^{(i)}} x_j^{(i)} \tag{16}$$

Given $\nabla \ell(\theta)_{\theta_j}$, we derive the stochastic gradient ascent update rule as follows:

$$\theta_j \leftarrow \theta_j + \alpha \times \nabla \ell(\theta)_{\theta_j} \tag{17}$$

$$\theta_j \leftarrow \theta_j + \alpha \times (y^{(i)} - e^{\theta^T x^{(i)}}) x_j^{(i)}. \tag{18}$$

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

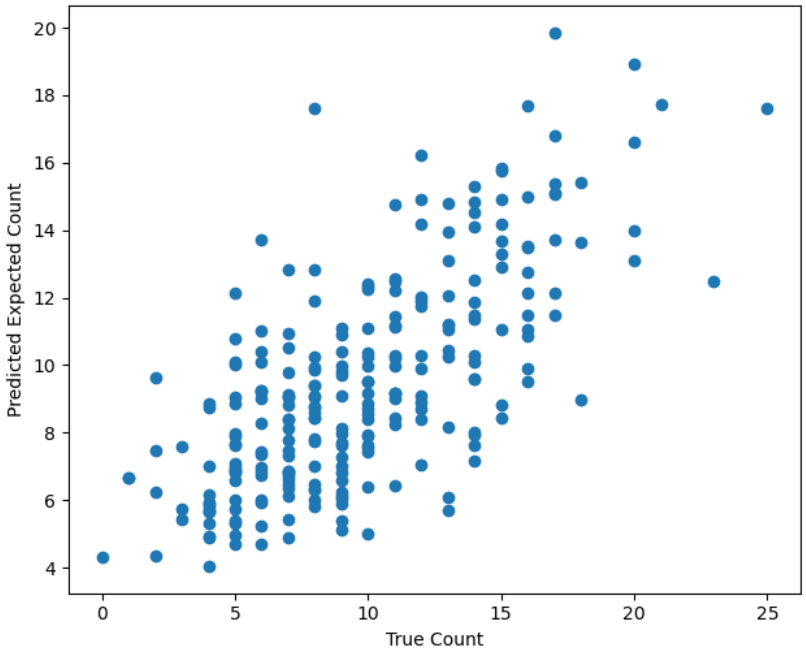
Consider a website that wants to predict its daily traffic. The website owners have collected a dataset of past traffic to their website, along with some features which they think are useful in predicting the number of visitors per day. The dataset is split into train/valid sets and the starter code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



2. **[15 points] Convexity of Generalized Linear Models**

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).

2(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: As per the hint, first observe:

$$\frac{\partial}{\partial\eta}\int p(y;\eta)dy = \int \frac{\partial}{\partial\eta}p(y;\eta)dy \tag{19}$$

$$= \int [0 \times \cdots + b(y)\exp(\eta y - a(\eta)) \times (y - \frac{\partial}{\partial\eta}a(\eta))]dy \tag{20}$$

$$= \int p(y;\eta) \times (y - \frac{\partial}{\partial\eta}a(\eta))dy \tag{21}$$

$$\frac{\partial}{\partial\eta}\int p(y;\eta)dy = \int yp(y;\eta)dy - \int p(y;\eta)\frac{\partial}{\partial\eta}a(\eta)dy \tag{22}$$

Further observe that:

- $\mathbb{E}[Y;\eta] = \int yp(y;\eta)dy$ by definition of expectation;
- $\frac{\partial}{\partial\eta}\int p(y;\eta)dy = 0$ since $\int p(y;\eta)dy = 1$ as $p(y;\eta)$ defines a probability distribution; and
- $\int p(y;\eta)\frac{\partial}{\partial\eta}a(\eta)dy = \frac{\partial}{\partial\eta}a(\eta)\int p(y;\eta)dy$ as $\frac{\partial}{\partial\eta}a(\eta)$ is a constant with respect to y.

Using these observations, we can simplify Equation 22 as follows:

$$0 = \mathbb{E}[Y;\eta] - \frac{\partial}{\partial\eta}a(\eta)\int p(y;\eta)dy$$

$$\mathbb{E}[Y;\eta] = \frac{\partial}{\partial\eta}a(\eta) \times 1$$

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

thereby demonstrating that which we wished to show ■.

2(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: As per the hint, we begin the derivation by observing the result of part (a):

$$\frac{\partial}{\partial \eta} a(\eta) = \mathbb{E}[Y;\eta] \tag{23}$$

$$\frac{\partial}{\partial \eta}\left(\frac{\partial}{\partial \eta} a(\eta)\right) = \frac{\partial}{\partial \eta}\mathbb{E}[Y;\eta] \tag{24}$$

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = \frac{\partial}{\partial \eta}\int y p(y;\eta) dy \tag{25}$$

$$= \int \frac{\partial}{\partial \eta} y p(y;\eta) dy \tag{26}$$

$$= \int y \times b(y) \exp(\eta y - a(\eta)) \times (y - \frac{\partial}{\partial \eta} a(\eta)) dy \tag{27}$$

$$= \int y p(y;\eta) \times (y - \frac{\partial}{\partial \eta} a(\eta)) dy \tag{28}$$

$$= \int y^2 p(y;\eta) dy - \int y p(y;\eta) \frac{\partial}{\partial \eta} a(\eta) dy \tag{29}$$

$$= \int y^2 p(y;\eta) dy - \frac{\partial}{\partial \eta} a(\eta) \int y p(y;\eta) \tag{30}$$

$$= \mathbb{E}[Y^2;\eta] - \mathbb{E}[Y;\eta] \times \mathbb{E}[Y;\eta] \tag{31}$$

$$\frac{\partial^2}{\partial \eta^2} a(\eta) = \mathbb{E}[Y^2;\eta] - (\mathbb{E}[Y;\eta])^2 = \text{Var}(Y;\eta) \tag{32}$$

thereby demonstrating that which we wished to show ■.

2(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:

First, we provide the loss function $\ell(\theta)$ as a function of θ:

$$\ell(\theta) = -\sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}; \theta) \tag{33}$$

$$= -\sum_{i=1}^{n} \log(b(y^{(i)}) \exp(\eta y^{(i)} - a(\eta))) \tag{34}$$

$$= -\sum_{i=1}^{n} \log(b(y^{(i)}) \exp(\theta^T x^{(i)} y^{(i)} - a(\theta^T x^{(i)}))) \tag{35}$$

$$= -\sum_{i=1}^{n} \log b(y^{(i)}) + \theta^T x^{(i)} y^{(i)} - a(\theta^T x^{(i)}). \tag{36}$$

To compute the Hessian of $\ell(\theta)$:

- We first compute the gradient of $\ell(\theta)$ w.r.t. θ:

$$\nabla_\theta \ell(\theta) = -\sum_{i=1}^{n} x^{(i)} y^{(i)} - \frac{\partial}{\partial \theta} a(\theta^T x^{(i)}) \times x^{(i)}. \tag{37}$$

- Then, we differentiate the gradient once more w.r.t θ to obtain the Hessian:

$$\mathbf{H}_\theta[\ell(\theta)] = -\sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} a(\theta^T x^{(i)}) \times x^{(i)} x^{(i)T} \tag{38}$$

$$= \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} a(\theta^T x^{(i)}) \times x^{(i)} x^{(i)T} \tag{39}$$

$$= \sum_{i=1}^{n} \mathsf{Var}(Y \mid X; \theta) \times x^{(i)} x^{(i)T}. \tag{40}$$

where Equation 40 follows from noting that: (1) for a fixed $x^{(i)}$, evaluating $\frac{\partial^2}{\partial \theta^2}$ is equivalent to evaluating $\frac{\partial^2}{\partial \eta^2}$; and (2) $\frac{\partial^2}{\partial \eta^2} a(\eta) = \mathsf{Var}(Y \mid X; \theta)$ as shown in part (b).

Finally, to observe that $\mathbf{H}_\theta[\ell(\theta)]$ is PSD, consider any $z \in \mathbb{R}^d$.

$$z^T \mathbf{H}_\theta[\ell(\theta)] z = \sum_{i=1}^{n} \mathsf{Var}(Y \mid X; \theta) \times (z^T x^{(i)})^2 = \sum_{i=1}^{n} \mathsf{Var}(Y \mid X; \theta) \times (\hat{x}^{(i)})^2.$$

Since the variance of any probability distribution as well as $(\hat{x}^{(i)})^2$ for any $\hat{x} \in \mathbb{R}^d$ (and associated $x \in \mathbb{R}^d$) are both non-negative, it follows that the Hessian is PSD. Thus, the NLL loss of GLM is convex w.r.t. the model parameters, i.e., θ ■.

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.

3. **[25 points] Linear regression: linear in what?**

In the first two lectures, you have seen how to fit a linear function of the data for the regression problem. In this question, we will see how linear regression can be used to fit non-linear functions of the data using feature maps. We will also explore some of its limitations, for which future lectures will discuss fixes.

3(a) [5 points] **Learning degree-3 polynomials of the input**

Suppose we have a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ where $x^{(i)}, y^{(i)} \in \mathbb{R}$. We would like to fit a third degree polynomial $h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x^1 + \theta_0$ to the dataset. The key observation here is that the function $h_\theta(x)$ is still linear in the unknown parameter θ, even though it's not linear in the input x. This allows us to convert the problem into a linear regression problem as follows.

Let $\phi : \mathbb{R} \to \mathbb{R}^4$ be a function that transforms the original input x to a 4-dimensional vector defined as

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix} \in \mathbb{R}^4 \tag{41}$$

Let $\hat{x} \in \mathbb{R}^4$ be a shorthand for $\phi(x)$, and let $\hat{x}^{(i)} \triangleq \phi(x^{(i)})$ be the transformed input in the training dataset. We construct a new dataset $\{(\phi(x^{(i)}), y^{(i)})\}_{i=1}^n = \{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^n$ by replacing the original inputs $x^{(i)}$'s by $\hat{x}^{(i)}$'s. We see that fitting $h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x^1 + \theta_0$ to the old dataset is equivalent to fitting a linear function $h_\theta(\hat{x}) = \theta_3 \hat{x}_3 + \theta_2 \hat{x}_2 + \theta_1 \hat{x}_1 + \theta_0$ to the new dataset because

$$h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x^1 + \theta_0 = \theta_3 \phi(x)_3 + \theta_2 \phi(x)_2 + \theta_1 \phi(x)_1 + \theta_0 = \theta^T \hat{x} \tag{42}$$

In other words, we can use linear regression on the new dataset to find parameters $\theta_0, \ldots, \theta_3$. Please write down 1) the objective function $J(\theta)$ of the linear regression problem on the new dataset $\{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^n$ and 2) the update rule of the batch gradient descent algorithm for linear regression on the dataset $\{(\hat{x}^{(i)}, y^{(i)})\}_{i=1}^n$.

Terminology: In machine learning, ϕ is often called the feature map which maps the original input x to a new set of variables. To distinguish between these two sets of variables, we will call x the input **attributes**, and call $\phi(x)$ the **features**. (Unfortunately, different authors use different terms to describe these two things. In this course, we will do our best to follow the above convention consistently.)

Answer:

1. The objective function $J(\theta)$ of the linear regression problem on the new dataset may be specified as follows:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (\theta^T \hat{x}^{(i)} - y^{(i)})^2. \tag{43}$$

2. The update rule of the batch gradient descent algorithm for linear regression on the new dataset may be specified as follows:

$$\theta \leftarrow \theta - \alpha \sum_{i=1}^{n} (\theta^T \hat{x}^{(i)} - y^{(i)}) \hat{x}^{(i)}. \tag{44}$$

Both equations are adaptations of formulas provided on pages 9 and 11, respectively, of the course reader, with $\theta^T \hat{x}$ and $\hat{x}$ being substituted for $h_\theta(x)$ and x as appropriate and with some term re-organization and associated sign changes for Equation 44 specifically.

3(b) [5 points] **Coding question: degree-3 polynomial regression**

For this sub-question question, we will use the dataset provided in the following files:

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

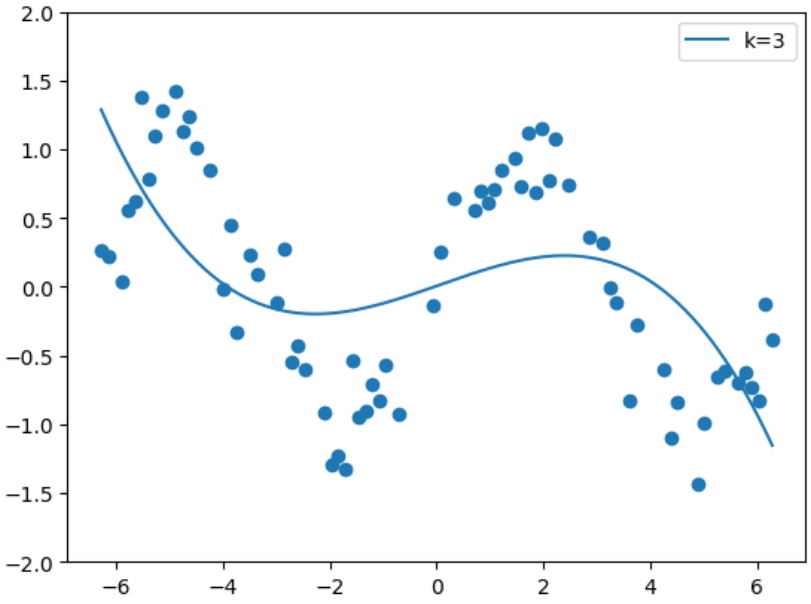
Each file contains two columns: x and y. In the terminology described in the introduction, x is the attribute (in this case one dimensional) and y is the output label.

Using the formulation of the previous sub-question, implement linear regression with **normal equations** using the feature map of degree-3 polynomials. Use the starter code provided in `src/featuremaps/featuremap.py` to implement the algorithm.

Create a scatter plot of the training data, and plot the learnt hypothesis as a smooth curve over it. Submit the plot in the writeup as the solution for this problem.

Remark: Suppose $\widehat{X}$ is the design matrix of the transformed dataset. You may sometimes encounter a non-invertible matrix $\widehat{X}^T\widehat{X}$. For a numerically stable code implementation, always use `np.linalg.solve` to obtain the parameters directly, rather than explicitly calculating the inverse and then multiplying it with $\widehat{X}^T y$.

Answer:



3(c) [5 points] **Coding question: degree-k polynomial regression**

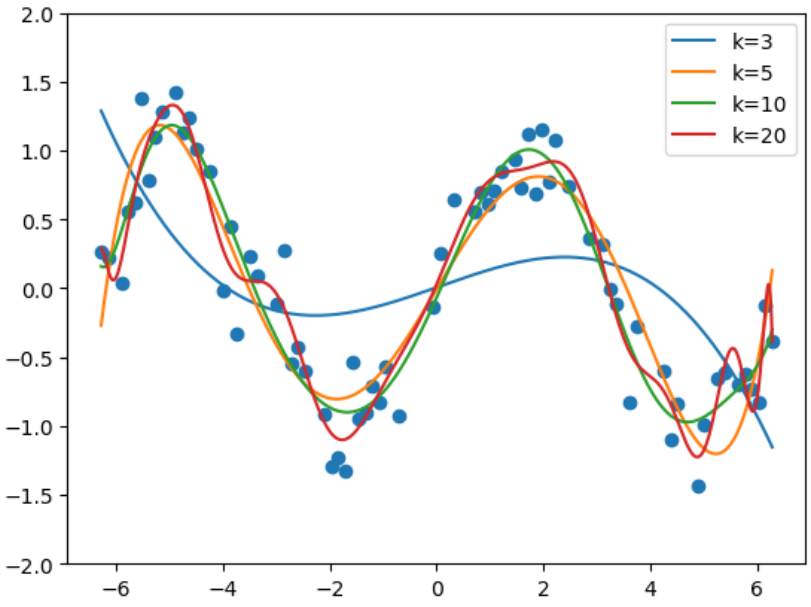
Now we extend the idea above to degree-k polynomials by considering $\phi : \mathbb{R} \to \mathbb{R}^{k+1}$ to be

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \in \mathbb{R}^{k+1} \tag{45}$$

Follow the same procedure as the previous sub-question, and implement the algorithm with $k = 3, 5, 10, 20$. Create a similar plot as in the previous sub-question, and include the hypothesis curves for each value of k with a different color. Include a legend in the plot to indicate which color is for which value of k.

Submit the plot in the writeup as the solution for this sub-problem. Observe how the fitting of the training dataset changes as k increases. Briefly comment on your observations in the plot.

Answer:



Observation: As k increases, the fitted regression more closely follows the distribution of the training data. At high values of k (i.e., $k = 20$), the fitted regression's curve becomes extremely close to the training data, likely indicating poor generalization to future data.

3(d) [5 points] **Coding question: other feature maps**

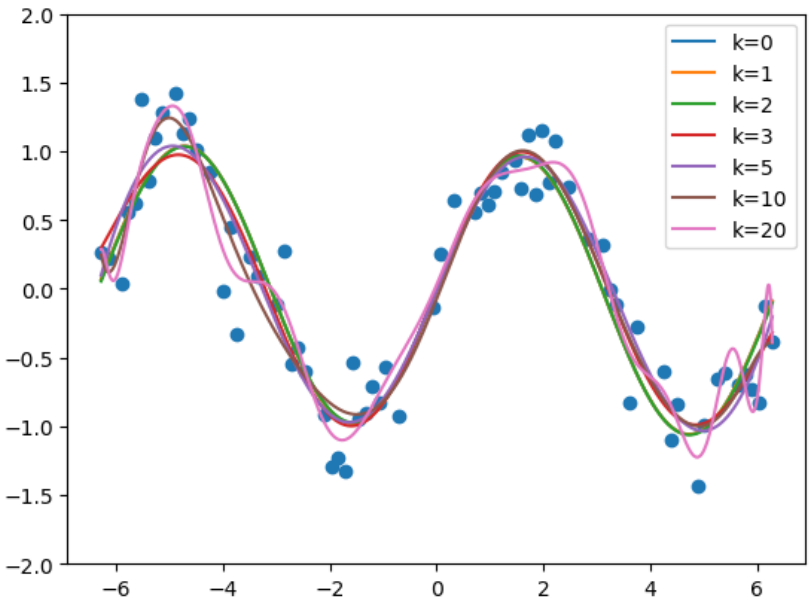
You may have observed that it requires a relatively high degree k to fit the given training data, and this is because the dataset cannot be explained (i.e., approximated) very well by low-degree polynomials. By visualizing the data, you may have realized that y can be approximated well by a sine wave. In fact, we generated the data by sampling from $y = \sin(x) + \xi$, where ξ is noise with Gaussian distribution. Please update the feature map ϕ to include a sine transformation as follows:

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \\ \sin(x) \end{bmatrix} \in \mathbb{R}^{k+2} \tag{46}$$

With the updated feature map, train different models for values of $k = 0, 1, 2, 3, 5, 10, 20$, and plot the resulting hypothesis curves over the data as before.

Submit the plot as a solution to this sub-problem. Compare the fitted models with the previous sub-question, and briefly comment about noticeable differences in the fit with this feature map.

Answer:



Observation: When comparing models to those generated for part (c), models with a sine feature and polynomial features up to degree k appear to fit the training data better as compared to models with only polynomial features up to the same degree k. Also, as k increases, the fit of models with the sine feature appear to improve less (as compared to the polynomial-only models from part (c)), likely due to the fact that the sine feature already captures much of the training data's distribution, and so there is less residual model error to capture with additional polynomial terms.

3(e) [5 points] **Overfitting with expressive models and small data**

For the rest of the problem, we will consider a small dataset (a random subset of the dataset you have been using so far) with much fewer examples, provided in the following file:

`src/featuremaps/small.csv`

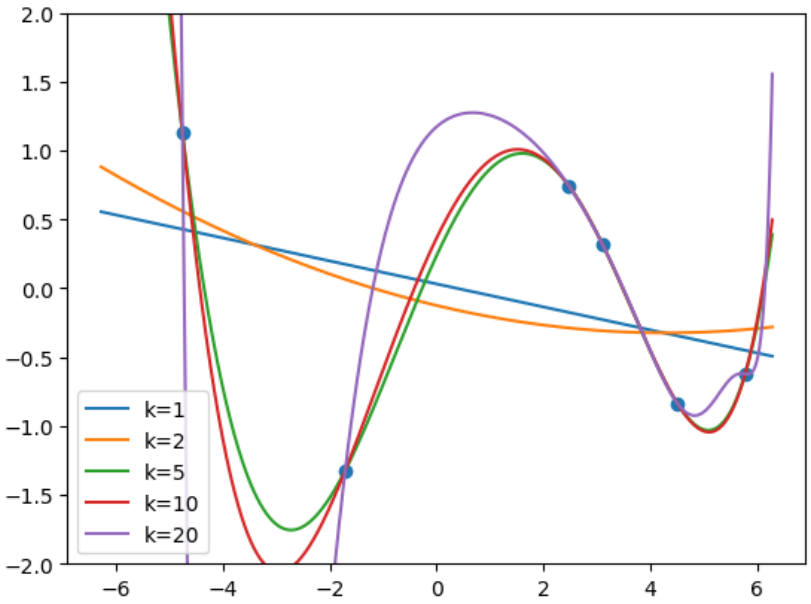
We will be exploring what happens when the number of features start becoming bigger than the number of examples in the training set. Run your algorithm on this small dataset using the following feature map

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^k \end{bmatrix} \in \mathbb{R}^{k+1} \tag{47}$$

with $k = 1, 2, 5, 10, 20$.

Create a plot of the various hypothesis curves (just like previous sub-questions). Observe how the fitting of the training dataset changes as k increases. Submit the plot in the writeup and comment on what you observe.

Answer:



Observation: When k is low ($k = 1, 2$), the fitted model curves do not perfectly intercept all of the training dataset points, but the general trend of the points appears to be captured. However, when k is high ($k \geq 5$), while the fitted model curves perfectly intercept all of the training dataset points, the fitted models have a highly erratic (i.e., frequently switching between increasing and decreasing), nonlinear shape—suggesting that these models are overfit to the training data and will generalize poorly when applied to new data with a predominately linear or low polynomial-degree trend.

Remark: The phenomenon you observe where the models start to fit the training dataset very well, but suddenly "goes wild" is due to what is called *overfitting*. The intuition to have for now is that, when the amount of data you have is small relative to the expressive capacity of the family of possible models (that is, the hypothesis class, which, in this case, is the family of all degree k polynomials), it results in overfitting.

Loosely speaking, the set of hypothesis function is "very flexible" and can be easily forced to pass through all your data points especially in unnatural ways. In other words, the model explains the noises in the training dataset, which shouldn't be explained in the first place. This hurts the predictive power of the model on test examples. We will describe overfitting in more detail in future lectures when we cover learning theory and bias-variance tradeoffs.

4. **[25 points] Learning Imbalanced dataset**

In this problem, we study how to learn a classifier from an imbalanced dataset, where the marginal distribution of the classes/labels are imbalanced. Imbalanced datasets are ubiquitous in real-world applications. For example, in the spam detection problem, the training dataset usually has only a small fraction of spam emails (positive examples) but a large fraction of ordinary emails (negative examples). For simplicity, we consider binary classification problem where the labels are in $\{0, 1\}$ and the number of positive examples is much smaller than the number of negative examples.

4(a) [10 points] **Evaluation metrics**

In this sub-question, we discuss the common evaluation metrics for imbalanced dataset. Suppose we have a validation dataset and for some $\rho \in (0, 1)$, we assume that ρ fraction of the validation examples are positive examples (with label 1), and $1 - \rho$ fraction of them are negative examples (with label 0).

Define the accuracy as

$$A \triangleq \frac{\#\text{examples that are predicted correctly by the classifier}}{\#\text{examples}}$$

(i) (3 point) Show that for any dataset with ρ fraction of positive examples and $1 - \rho$ fraction of negative examples, there exists a (trivial) classifier with accuracy at least $1 - \rho$.

The statement above suggests that the accuracy is not an ideal evaluation metric when ρ is close to 0. E.g., imagine that for spam detection ρ can be smaller than 1%. The statement suggests there is a trivial classifier that gets more than 99% accuracy. This could be misleading — 99% seems to be almost perfect, but actually you don't need to learn anything from the dataset to achieve it.

Therefore, for imbalanced dataset, we need more informative evaluation metrics. We define the number of true positive, true negative, false positive, false negative examples as

$$TP \triangleq \#\text{ positive examples with a correct (positive) prediction}$$
$$TN \triangleq \#\text{ negative examples with a correct (negative) prediction}$$
$$FP \triangleq \#\text{ negative examples with a incorrect (positive) prediction}$$
$$FN \triangleq \#\text{ positive examples with a incorrect (negative) prediction}$$

Define the accuracy of positive examples as

$$A_1 \triangleq \frac{TP}{TP + FN} = \frac{\#\text{ positive examples with a correct (positive) prediction}}{\#\text{ positive examples}}$$

Define the accuracy of negative examples as

$$A_0 \triangleq \frac{TN}{TN + FP} = \frac{\#\text{ negative examples with a correct (negative) prediction}}{\#\text{ negative examples}}$$

We define the balanced accuracy as

$$\overline{A} \triangleq \frac{1}{2}(A_0 + A_1) \tag{48}$$

With these notations, we can verify that the accuracy is equal to $A = \frac{TP+TN}{TP+TN+FP+FN}$.

(ii) (4 point) Show that

$$\rho = \frac{TP + FN}{TP + TN + FP + FN}$$

and

$$A = \rho \cdot A_1 + (1 - \rho)A_0 \tag{49}$$

Comparing equation (48) and (49), we can see that the accuracy and balanced accuracy are both linear combination of A_0 and A_1 but with different weighting.

(iii) (3 point) Show that the trivial classifier you constructed for part (i) has balanced accuracy 50%.

Partly because of (iii), the balanced accuracy $\overline{A}$ is often a preferable evaluation metric than the accuracy A. Sometimes people also report the accuracies for the two classes (A_0 and A_1) to demonstrate the performance for each class.

Answer:

(i) Consider the classifier which does not train and predicts the negative label (0) for all inputs during evaluation. In any dataset with ρ fraction of positive examples and $1-\rho$ fraction of negative examples, this classifier will produce the correct label for all negative examples. Since there are $1-\rho$ fraction of negative examples, this trivial classifier achieves an accuracy of at least $1-\rho$, as required ■.

(ii) (Part 1) To see that:

$$\rho = \frac{TP+FN}{TP+TN+FP+FN}$$

let us first define ρ as follows:

$$\rho = \frac{\#\text{ of examples with a positive label}}{\#\text{ total number of examples}}$$

where this definition follows from ρ being the fraction of positive examples in the dataset. For every example with a positive label, an arbitrary classifier will either correctly or incorrectly predict a positive label. Therefore, # of examples with a positive label = # of positive examples with a correct prediction + # of positive examples with an incorrect prediction. Equivalently, # of examples with a positive label $= TP+FN$.

Further observe that the # total number of examples = # of examples with a positive label + # of examples with a negative label. Analogous to the positive label case, an arbitrary classifier will either correctly or incorrectly predict a negative label. Therefore, # of examples with a negative label = # of negative examples with a correct prediction + # of negative examples with an incorrect prediction. Equivalently, # of examples with a negative label $= TN+FP$.

With these definitions, observe that # total number of examples = # of examples with a positive label + # of examples with a negative label $= TP+FN+TN+FP$. Given this, we can encode ρ as follows:

$$\rho = \frac{\#\text{ of examples with a positive label}}{\#\text{ total number of examples}} = \frac{TP+FN}{TP+TN+FP+FN}$$

verifying Equation 48 as desired ■.

(Part 2) To see that $A = \rho \cdot A_1 + (1-\rho)\cdot A_0$, we can expand the RHS of this equation using the definitions of A_0, A_1, ρ determined thus far:

$$
\begin{aligned}
&\rho \cdot A_1 + (1-\rho) \cdot A_0 \\
&= \frac{TP+FN}{TP+TN+FP+FN} \cdot \frac{TP}{TP+FN} + (1 - \frac{TP+FN}{TP+TN+FP+FN}) \cdot \frac{TN}{TN+FP} \\
&= \frac{TP}{TP+TN+FP+FN} + \frac{TN+FP}{TP+TN+FP+FN} \cdot \frac{TN}{TN+FP} \\
&= \frac{TP}{TP+TN+FP+FN} + \frac{TN}{TP+TN+FP+FN} \\
&= \frac{TP+TN}{TP+TN+FP+FN} = A
\end{aligned}
$$

verifying Equation 49 as desired ■.

(iii) To see that the trivial classifier defined in (i) will achieve a balanced accuracy $\overline{A}$ of 50%, note that said classifier will predict the correct label for all negative label examples and will predict the incorrect label for all positive label examples. As such, $A_1 = 0\%$ and $A_0 = 100\%$. Then, the classifier's balanced accuracy can be computed as (per Equation 48):

$$\overline{A} = \frac{1}{2}(100\% + 0\%) = 50\%$$

demonstrating that which we wished to show ■.

4(b) [5 points] **Coding problem: vanilla logistic regression**

First, we use the vanilla logistic regression to learn an imbalanced dataset. For the rest of the question, we will use the dataset and starter code provided in the following files:

- `src/imbalanced/{train,validation}.csv`
- `src/imbalanced/imbalanced.py`

Each file contains n examples, one example $(x^{(i)}, y^{(i)})$ per row. x is two-dimensional, i.e., the i-th row contains columns $x_1^{(i)} \in \mathbb{R}$, $x_2^{(i)} \in \mathbb{R}$, and $y^{(i)} \in \{0, 1\}$. Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$ be our training dataset. $\mathcal{D}$ has ρn examples with label 1 and $(1-\rho)n$ with label 0. In the dataset we constructed, $\rho = 1/11$.

You will train a linear classifier $h_\theta(x)$ with average empirical loss for logistical regression, where $h_\theta(x) = g(\theta^T x), g(z) = 1/(1+e^{-z})$:

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

You can use the provided logistic regression implementation, or any standard logistic regression library to optimize the objective above. After obtaining the classifier, compute the classifier's accuracy (A), balanced accuracy ($\overline{A}$), accuracies for the two classes (A_0, A_1) on the validation dataset, and report them in the writeup. You are expected to observe that the minority class (positive class) has significantly lower accuracy than the majority class.

Create a plot to visualize the validation set with x_1 on the horizontal axis and x_2 on the vertical axis. Use different symbols for examples $x^{(i)}$ with true label $y^{(i)} = 1$ than those with $y^{(i)} = 0$. On the same figure, plot the decision boundary obtained by your model (i.e, line corresponding to model's predicted probability = 0.5) in red color. Include this plot in your writeup.

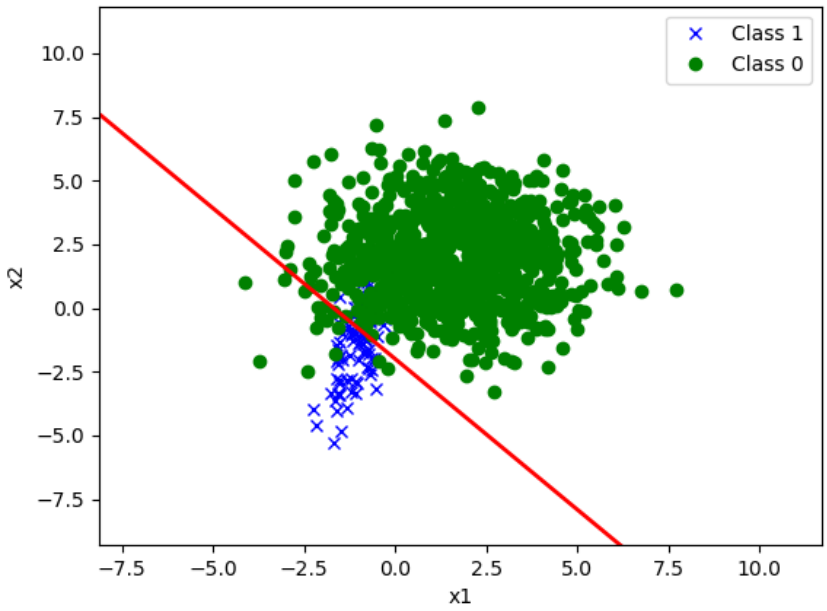
Answer:

Accuracy statistics:

- Overall accuracy (A): 94.73%
- Balanced accuracy ($\overline{A}$): 77.75%
- Negative class accuracy (A_0): 98.50%
- Positive class accuracy (A_1): 57.00%

Plot:

(See next page...)


Class 1
Class 0
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
x2
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
x1

4(c) [5 points] **Re-sampling/Re-weighting Logistic Regression**

The relatively low accuracy for the minority class and the resulting low balanced accuracy are undesirable for many applications. Various methods have been proposed to improve the accuracy for the minority class, and learning imbalanced datasets is an active open research direction. Here we introduce a simple and classical re-sampling/re-weighting technique that helps improve the balanced accuracy in most of the cases.

We observe that the logistic regression algorithm works well for the accuracy but not for the balanced accuracy. Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ denote the existing training dataset. We will create a new dataset $\mathcal{D}'$ such that the accuracy on $\mathcal{D}'$ is the same as the balanced accuracy on $\mathcal{D}$.

Assume $\rho < 1/2$ without loss of generality, and let $\kappa = \frac{\rho}{1-\rho}$. This means the number of positive examples is κ times the number of negative examples in $\mathcal{D}$. Assume for convenience $1/\kappa$ is an integer.[3] Let $\mathcal{D}'$ be the dataset that contain each negative example once and $1/\kappa$ repetitions of each positive example in $\mathcal{D}$. Then we will apply the logistic regression on the new dataset $\mathcal{D}'$.

Prove that for any classifier, the balanced accuracy on $\mathcal{D}$ is equal to the accuracy on $\mathcal{D}'$. Moreover, **show that** the average empirical loss for logistical regression on the dataset $\mathcal{D}'$ is equal to

$$J(\theta) = -\frac{1+\kappa}{2n} \sum_{i=1}^{n} w^{(i)} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right),$$

where $w^{(i)} = 1$ if $y^{(i)} = 0$ and $w^{(i)} = 1/\kappa$ if $y^{(i)} = 1$.

Observe effectively we are re-weighting the loss function for each example by some constant that depends on the frequency of the class it belongs to.

Answer:

(Part 1) Let Q' denote the analogous quantity for dataset $\mathcal{D}'$ as related to quantity Q for dataset D; e.g., the number of true positives for any classifier on $\mathcal{D}$ will be denoted TP whereas the number of true positives for the same classifier on $\mathcal{D}'$ will be denoted TP'.

Consider an arbitrary classifier C. Observe that $TP' = TP \times \frac{1}{\kappa}$ and $FN' = FN \times \frac{1}{\kappa}$ while $TN' = TN$ and $FP' = FP$. This follows from $\mathcal{D}'$ being created by copying the positive label examples of $\mathcal{D}$ $\frac{1}{\kappa}$ times each while copying each negative label example only once. We can now encode the accuracy achieved by C on $\mathcal{D}'$ as follows:

$$A(\mathcal{D}') = \frac{TP' + TN'}{TP' + TN' + FP' + FN'} \tag{50}$$

$$= \frac{\frac{TP}{\kappa} + TN}{\frac{TP}{\kappa} + TN + FP + \frac{FN}{\kappa}} \tag{51}$$

$$= \frac{TP \times \frac{1-\rho}{\rho} + TN}{TP \times \frac{1-\rho}{\rho} + TN + FP + FN \times \frac{1-\rho}{\rho}} \quad \text{(substitute } \kappa = \frac{\rho}{1-\rho}\text{)} \tag{52}$$

$$= \frac{TP(1-\rho) + TN\rho}{TP(1-\rho) + TN\rho + FP\rho + FN(1-\rho)} \quad \text{(multiply (52) by } 1 = \frac{\rho}{\rho}\text{)} \tag{53}$$

$$= \frac{TP(1-\rho) + TN\rho}{(1-\rho)(TP + FN) + \rho(TN + FP)} \tag{54}$$

[3] otherwise we can round up to the nearest integer with introducing a slight approximation.

From part (a)(ii), recall

$$\rho = \frac{TP + FN}{TP + TN + FP + FN}.$$

It follows that

$$1 - \rho = \frac{TN + FP}{TP + TN + FP + FN}.$$

Also observe $|\mathcal{D}| = TP + TN + FP + FN$. We can substitute these equalities into Equation 54 and simplify as follows:

$$A(\mathcal{D}') = \frac{TP \times \frac{TN+FP}{|\mathcal{D}|} + TN \times \frac{TP+FN}{|\mathcal{D}|}}{\frac{TN+FP}{|\mathcal{D}|} \times (TP + FN) + \frac{TP+FN}{|\mathcal{D}|} \times (TN + FP)} \tag{55}$$

$$= \frac{TP(TN + FP) + TN(TP + FN)}{(TN + FP)(TP + FN) + (TP + FN)(TN + FP)} \tag{56}$$

$$= \frac{TP(TN + FP) + TN(TP + FN)}{2 \times (TN + FP)(TP + FN)} \tag{57}$$

$$= \frac{1}{2} \times \left(\frac{TP(TN + FP)}{(TN + FP)(TP + FN)} + \frac{TN(TP + FN)}{(TN + FP)(TP + FN)} \right) \tag{58}$$

$$= \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{59}$$

$$= \frac{1}{2}(A_1(\mathcal{D}) + A_0(\mathcal{D})) \tag{60}$$

$$A(\mathcal{D}') = \overline{A}(\mathcal{D}) \tag{61}$$

thereby demonstrating that the balanced accuracy on $\mathcal{D}$ is equal to the accuracy on $\mathcal{D}'$ for any arbitrary classifier C ■.

(Part 2) Consider the average empirical loss for logistic regression on dataset $\mathcal{D}$; i.e.,

$$J_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right) \tag{62}$$

where $h_\theta(x) = g(\theta^T x), g(z) = 1/(1 + e^{-z})$. Let us denote summands of Equation 62 (i.e., the empirical loss on a single example $(x^{(i)}, y^{(i)})$) as Loss_i. For instance, we can re-write Equation 62 as follows:

$$J_{\mathcal{D}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \text{Loss}_i. \tag{63}$$

Now consider the average empirical loss for logistic regression on dataset $\mathcal{D}'$; i.e.,

$$J_{\mathcal{D}'}(\theta) = -\frac{1}{|\mathcal{D}'|} \sum_{i=1}^{|\mathcal{D}'|} \left(y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right) \tag{64}$$

$$= -\frac{1}{|\mathcal{D}'|} \sum_{i=1}^{|\mathcal{D}'|} \text{Loss}_i \tag{65}$$

Notice we can decompose the summation of Equation 65 into two constituent summations, one over dataset examples where $y^{(i)} = 1$ and one over dataset examples where $y^{(i)} = 0$:

$$J_{\mathcal{D}'}(\theta) = -\frac{1}{|\mathcal{D}'|}\left(\sum_{\substack{i=1 \\ y^{(i)}=1}}^{|\mathcal{D}'|} \text{Loss}_i + \sum_{\substack{i=1 \\ y^{(i)}=0}}^{|\mathcal{D}'|} \text{Loss}_i \right) \tag{66}$$

We can decompose the average empirical loss on dataset $\mathcal{D}$ similarly:

$$J_{\mathcal{D}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\substack{i=1 \\ y^{(i)}=1}}^{n} \text{Loss}_i + \sum_{\substack{i=1 \\ y^{(i)}=0}}^{n} \text{Loss}_i \right) \tag{67}$$

Observe that, since datasets $\mathcal{D}$ and $\mathcal{D}'$ have the same number of *identical* negative label examples, it must be true that:

$$\sum_{\substack{i=1 \\ y^{(i)}=0}}^{|\mathcal{D}'|} \text{Loss}_i = \sum_{\substack{i=1 \\ y^{(i)}=0}}^{n} 1 \times \text{Loss}_i \tag{68}$$

Analogously, since dataset $\mathcal{D}'$ has $\frac{1}{\kappa}$ identical copies of every positive label example in $\mathcal{D}$, it must also be true that:

$$\sum_{\substack{i=1 \\ y^{(i)}=1}}^{|\mathcal{D}'|} \text{Loss}_i = \sum_{\substack{i=1 \\ y^{(i)}=1}}^{n} \frac{1}{\kappa} \times \text{Loss}_i \tag{69}$$

We can substitute Equations 68 and 69 into Equation 66 to observe:

$$J_{\mathcal{D}'}(\theta) = -\frac{1}{|\mathcal{D}'|}\left(\sum_{\substack{i=1 \\ y^{(i)}=0}}^{n} 1 \times \text{Loss}_i + \sum_{\substack{i=1 \\ y^{(i)}=1}}^{n} \frac{1}{\kappa} \times \text{Loss}_i \right) \tag{70}$$

$$= -\frac{1}{|\mathcal{D}'|}\left(\sum_{i=1}^{n} w^{(i)} \times \text{Loss}_i \right) \tag{71}$$

where $w^{(i)} = 1$ if $y^{(i)} = 0$ and $w^{(i)} = 1/\kappa$ if $y^{(i)} = 1$.

Finally, observe that dataset $\mathcal{D}$ contains ρn positive label examples and $(1-\rho)n$ negative label examples. By construction, dataset $\mathcal{D}'$ therefore contains $\frac{1}{\kappa}\rho n$ positive label examples and

$(1-\rho)n$ negative label examples. Recall $\kappa = \frac{\rho}{1-\rho}$ and thus $1-\rho = \frac{\rho}{\kappa}$, $\rho = \kappa(1-\rho)$. Therefore:

$$|\mathcal{D}'| = \frac{1}{\kappa}\rho n + (1-\rho)n \tag{72}$$
$$= \frac{\rho n}{\kappa} + \frac{\rho n}{\kappa} = \frac{2\rho n}{\kappa} \tag{73}$$
$$= \frac{2n}{\kappa} \times \rho = \frac{2n}{\kappa} \times (\kappa(1-\rho)) \tag{74}$$
$$= 2n \times (1-\rho) = 2n \times (1-\kappa(1-\rho)) \tag{75}$$
$$= 2n \times (1-\kappa+\kappa\rho) = (1-\kappa+\kappa(\kappa(1-\rho))) \tag{76}$$
$$= 2n \times (1-\kappa+\kappa^2-\kappa^2\rho) = 2n \times (1-\kappa+\kappa^2-\kappa^2(\kappa(1-\rho)) \tag{77}$$
$$= 2n \times (1-\kappa+\kappa^2-\kappa^3+\kappa^3\rho) \tag{78}$$
$$= 2n \times \Big((1+\kappa^2+\cdots)-(\kappa+\kappa^3+\cdots)\Big) \tag{79}$$
$$= 2n \times \left(\frac{1}{1-\kappa^2} - \frac{\kappa}{1-\kappa^2}\right) \text{ (valid finite approximations as } \kappa < 1) \tag{80}$$
$$= 2n \times (\frac{1-\kappa}{1-\kappa^2}) \tag{81}$$
$$= 2n \times (\frac{1}{1+\kappa}) = \frac{2n}{1+\kappa}. \tag{82}$$

Substituting Equation 82 into Equation 71 and fully expressing Loss_i, we thus observe that:

$$J_{\mathcal{D}'}(\theta) = -\frac{1}{\frac{2n}{1+\kappa}}\left(\sum_{i=1}^{n} w^{(i)} \times \text{Loss}_i\right) \tag{83}$$
$$= -\frac{1+\kappa}{2n}\sum_{i=1}^{n} w^{(i)} \times \left(y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right) \tag{84}$$

where $w^{(i)} = 1$ if $y^{(i)} = 0$ and $w^{(i)} = 1/\kappa$ if $y^{(i)} = 1$, thereby demonstrating that which we wished to prove ■.

4(d) [5 points] **Coding problem: re-weighting minority class**

In `src/imbalanced/imbalanced.py`, implement the logistic regression algorithm on $\mathcal{D}'$. Compute and report the classifier's accuracy (A), balanced accuracy ($\overline{A}$), accuracies for the two classes (A_0, A_1) on the validation dataset. You are expected to see that the accuracy of minority class (class 1) improved significantly whereas that of the majority class dropped compared to vanilla logistic regression. However, the balanced accuracy is significantly greater than that of the vanilla logistic regression.

Create a plot to visualize the validation set with x_1 on the horizontal axis and x_2 on the vertical axis. Use different symbols for examples $x^{(i)}$ with true label $y^{(i)} = 1$ than those with $y^{(i)} = 0$. On the same figure, plot the decision boundary obtained by your model (i.e, line corresponding to model's predicted probability = 0.5) in red color. Include this plot in your writeup.

Answer:

Accuracy statistics:

- Overall accuracy (A): 89.91%
- Balanced accuracy ($\overline{A}$): 90.40%
- Negative class accuracy (A_0): 89.80%
- Positive class accuracy (A_1): 91.00%

Plot: