

Homework 2: Classification and Bias-Variance Trade-offs

Introduction

This homework is about classification and bias-variance trade-offs. In lecture we have primarily focused on binary classifiers trained to discriminate between two classes. In multiclass classification, we discriminate between three or more classes. Most of the material for Problem 1 and Problem 3, and all of the material for Problem 2 will be covered by the end of the Tuesday 2/8 lecture. The rest of the material will be covered by the end of the Thursday 2/10 lecture. We encourage you to read CS181 Textbook's Chapter 3 for more information on linear classification, gradient descent, classification in the discriminative setting (covers multiclass logistic regression and softmax), and classification in the generative setting. Read Chapter 2.8 for more information on the trade-offs between bias and variance.

As a general note, for classification problems we imagine that we have the input matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ (or perhaps they have been mapped to some basis Φ , without loss of generality) with outputs now “one-hot encoded.” This means that if there are K output classes, rather than representing the output label y as an integer $1, 2, \dots, K$, we represent \mathbf{y} as a “one-hot” vector of length K . A “one-hot” vector is defined as having every component equal to 0 except for a single component which has value equal to 1. For example, if there are $K = 7$ classes and a particular data point belongs to class 3, then the target vector for this data point would be $\mathbf{y} = [0, 0, 1, 0, 0, 0, 0]$. We will define C_1 to be the one-hot vector for the 1st class, C_2 for the 2nd class, etc. Thus, in the previous example $\mathbf{y} = C_3$. If there are K total classes, then the set of possible labels is $\{C_1 \dots C_K\} = \{C_k\}_{k=1}^K$. Throughout the assignment we will assume that each label $\mathbf{y} \in \{C_k\}_{k=1}^K$ unless otherwise specified. The most common exception is the case of binary classification ($K = 2$), in which case labels are the typical integers $y \in \{0, 1\}$.

In problems 1 and 3, you may use `numpy` or `scipy`, but not `scipy.optimize` or `sklearn`. Example code given is in Python 3.

Please type your solutions after the corresponding problems using this L^AT_EX template, and start each problem on a new page.

Please submit the **writeup PDF to the Gradescope assignment ‘HW2’**. Remember to assign pages for each question. **You must include your plots in your writeup PDF**. The supplemental files will only be checked in special cases, e.g. honor code issues, etc.

Please submit your **L^AT_EX file and code files to the Gradescope assignment ‘HW2 - Supplemental’**.

Problem 1 (Exploring Bias and Variance, 10 pts)

In this problem, we will explore the bias and variance of a few different model classes when it comes to logistic regression.

Consider the true data generating process $y \sim \text{Bern}(f(x))$, $f(x) = 0.4 \times \sin(1.2x) + 0.5$, where $x \in [-3, 3]$, and $y \in \{0, 1\}$. Recall that for a given x , bias and variance are defined in terms of expectations *over randomly drawn datasets D* from this underlying data distribution:

$$\begin{aligned}\text{Bias}[\hat{f}(x)] &= \mathbb{E}_D[\hat{f}(x)] - f(x) \\ \text{Variance}[\hat{f}(x)] &= \mathbb{E}_D[(\hat{f}(x) - \mathbb{E}_D[\hat{f}(x)])^2]\end{aligned}$$

Here, $\hat{f}(x)$ is our estimator (learned through logistic regression on a given dataset D). We will directly explore the bias-variance trade-off by drawing multiple such datasets and fitting different logistic regression models to each. Remember that we, the modelers, do not usually see the true data distribution. Knowledge of the true $f(x)$ is only exposed in this problem to (1) make possible the simulation of drawing multiple datasets, and (2) to serve as a pedagogical tool in allowing verification of the true bias.

1. Consider the three bases $\phi_1(x) = [1, x]$, $\phi_2(x) = [1, x, x^2]$, $\phi_3(x) = [1, x, x^2, x^3, x^4, x^5]$. For each of these bases, generate 10 datasets of size $N = 30$ using the starter code provided, and fit a logistic regression model using $\text{sigmoid}(w^T \phi(x))$ to each dataset by using gradient descent to minimize the negative log likelihood. This means you will be running gradient descent 10 times for each basis, once for each dataset. Note that the classes are represented with 0's and 1's.

Use random starting values of w , $\eta = 0.001$, take 10,000 update steps for each gradient descent run, and make sure to average the gradient over the data points (for each step). These parameters, while not perfect, will ensure your code runs in a reasonable amount of time. The emphasis of this problem is on capturing the bias-variance trade-off, so don't worry about attaining perfect precision in the gradient descent as long as this trade-off is captured in the final models.

Note: Overflow RuntimeWarnings due to `np.exp` should be safe to ignore, if any. Also, to reduce stress from randomness in students' solutions (due to randomized weight initialization differences), in line 109 of the `T2_P1.py` starter code, we call `np.random.seed(1738)` to set a deterministic random seed. Please do not change this! In addition, please do not change the randomized weight initialization code in lines 42 – 46.

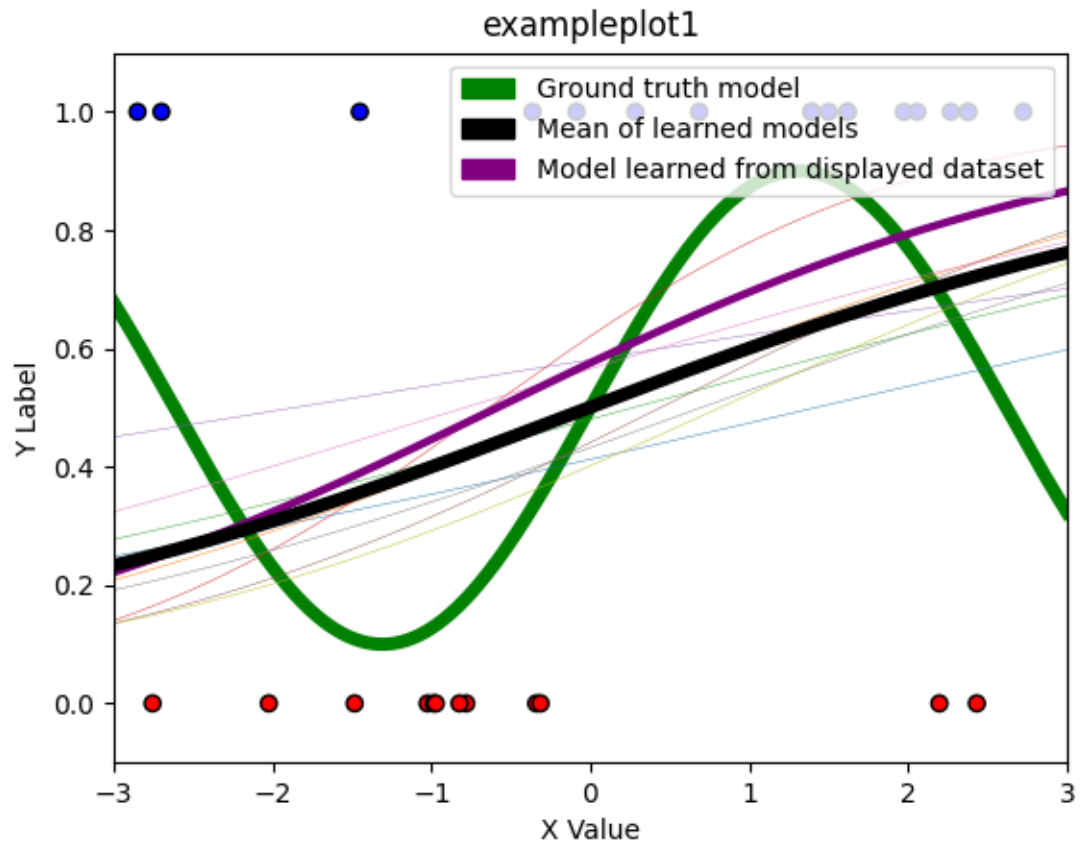
2. Create three plots, one for each basis. Starter code is available which you may modify. By default, each plot displays three types of functions: (1) the true data-generating distribution $f(x)$ (the probability that $y = 1$ for different x). (2) all 10 of the prediction functions learned from each randomly drawn dataset, and (3) the mean of the 10 prediction functions. Moreover, each plot also displays 1 of the randomly generated datasets and highlights the corresponding prediction function learned by this dataset.
3. How are bias and variance reflected in the 3 types of curves on the graphs? How do the fits of the individual and mean prediction functions change? Keeping in mind that none of the model classes match the true generating process exactly, discuss the extent to which each of the bases approximates the true process.

Note: In this problem, we are not interested in whether the model is more biased for certain inputs x compared to other inputs x' . We are interested in the overall bias and variance of $\hat{f}(x)$ across the different basis choices. In other words, we want to investigate how the bias between $\hat{f}(x)$ and the ground truth as well as the variance of $\hat{f}(x)$ will be different over different basis choices.

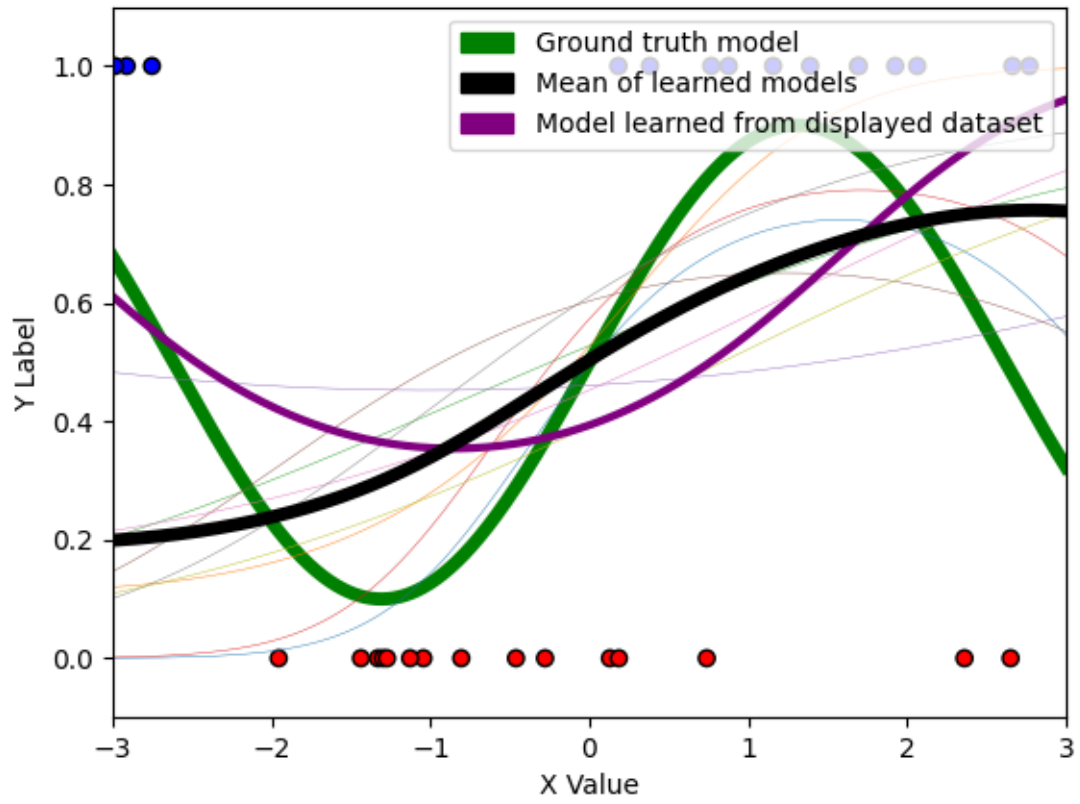
4. If we were to increase the size of each dataset drawn from $N = 30$ to a larger number, how would the variance change? The bias? Why might this be the case?

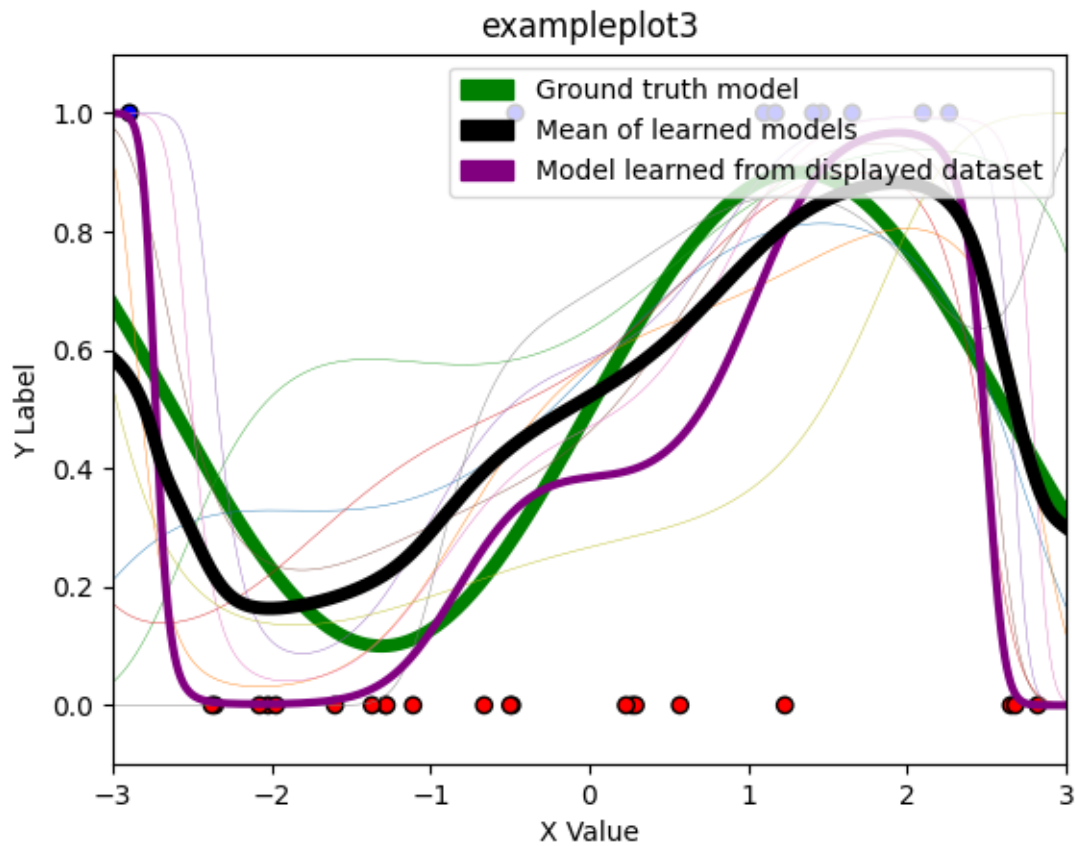
Solution

- 1.1 See source code.
- 1.2

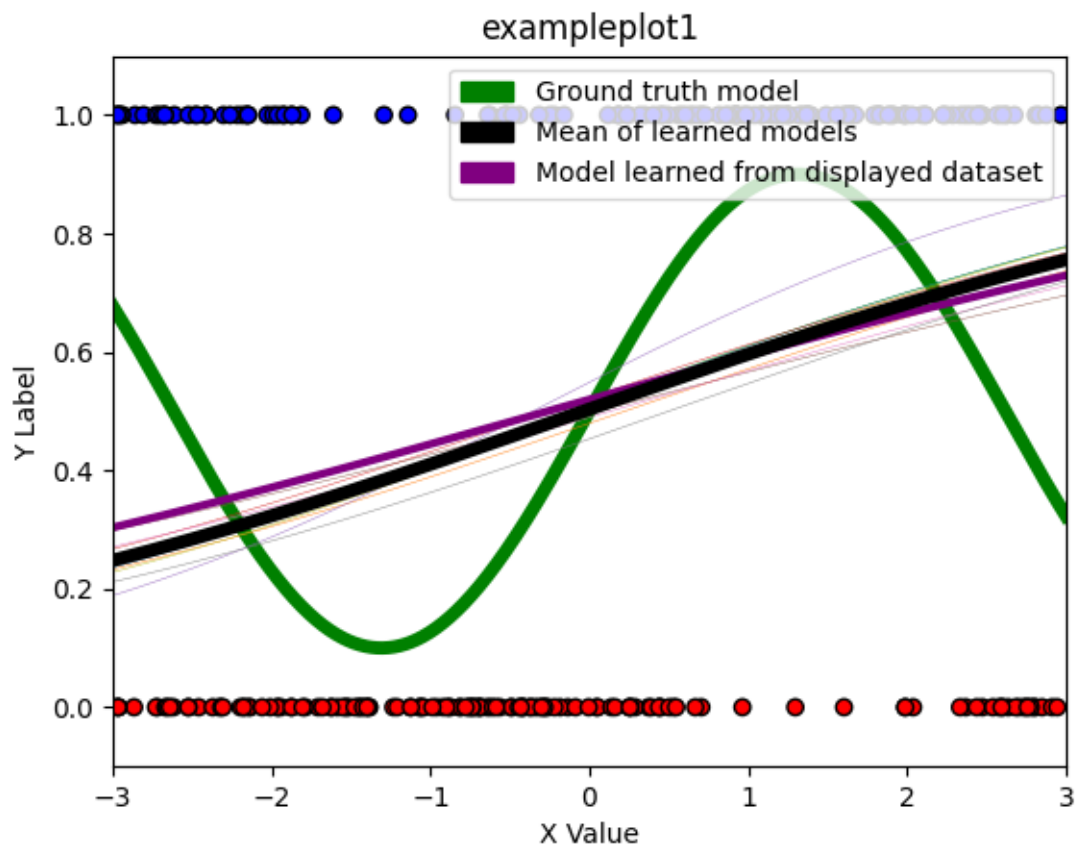


exampleplot2



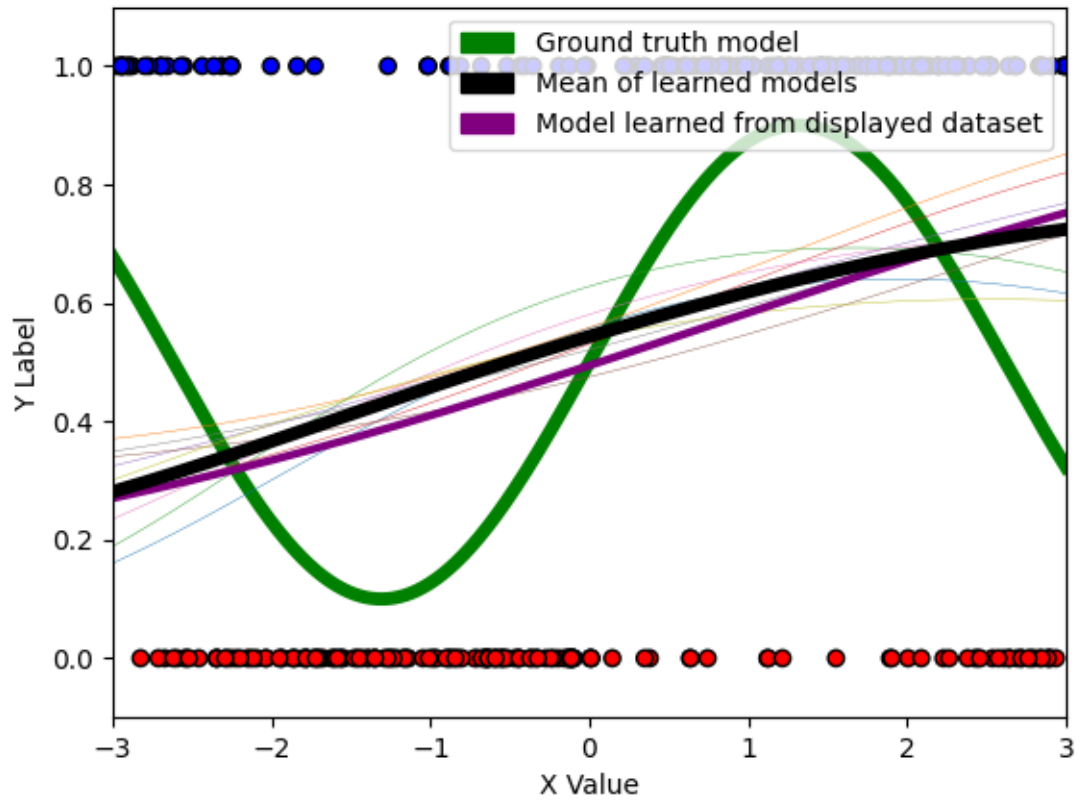


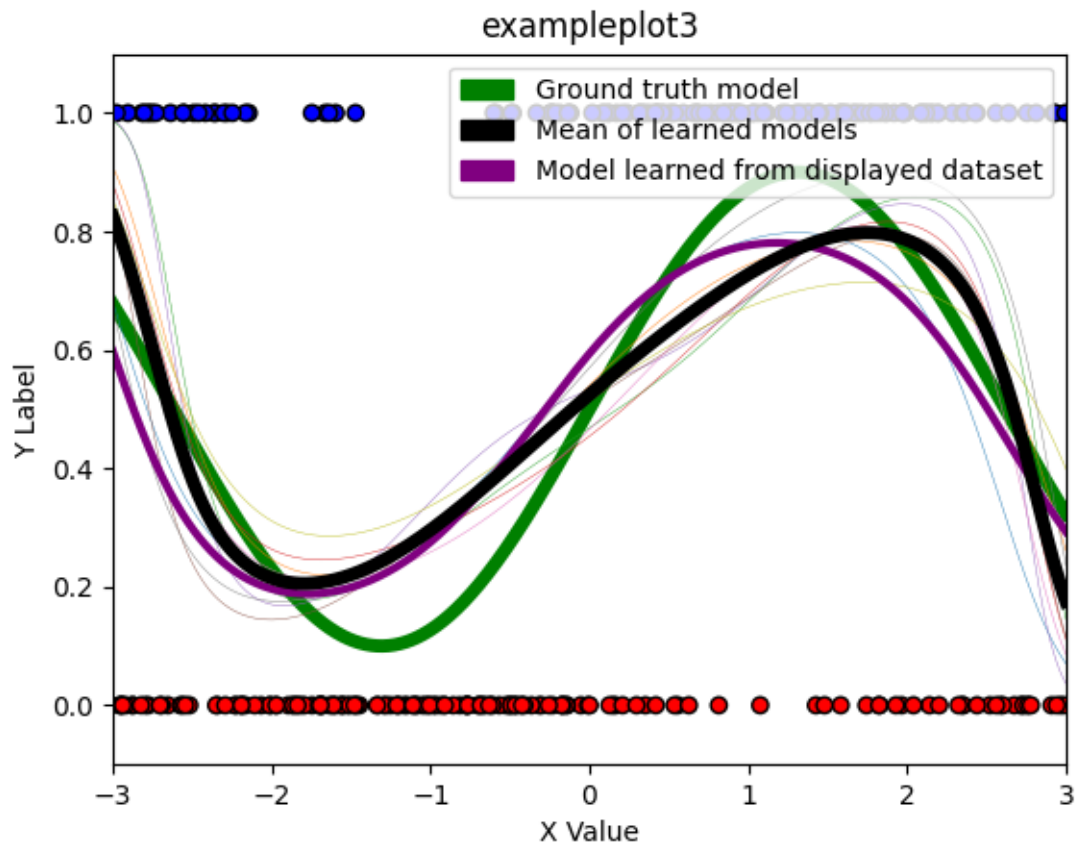
3. As the bases become more complex, the bias decreases as the curves adhere more closely to the sine function, but the variance increases. Note that if one point were to be moved, the regression would fit a completely different polynomial to go through that new location. In example plot 3, the final polynomial the model settles on (purple curve) exhibits high variance: notice how it continues to increase even after the sine function starts declining at $x=1$, because the purple curve is heavily influenced by the two training points around $x=2$, $y=1$. If the dataset were randomly generated again, the purple curve might exhibit very different behavior in this vicinity.



4.

exampleplot2





As n increases, bias stays the same but variance decreases. This is because the model is already approximating the average behavior of the underlying distribution even at low dataset sizes, but with more data points in the sample, it can better follow the trends and a single random outlier will have less of an impact and not drag the model as far in an idiosyncratic direction. More data points thus counteract the overfitting nature of highly complex bases. (Note that if the basis has more dimensions than the number of data points, the system will be overdetermined and the model can get zero error even without learning the true relationship between the predictors and the outcomes.)

Problem 2 (Maximum likelihood in classification, 15pts)

Consider now a generative K -class model. We adopt class prior $p(\mathbf{y} = C_k; \boldsymbol{\pi}) = \pi_k$ for all $k \in \{1, \dots, K\}$ (where π_k is a parameter of the prior). Let $p(\mathbf{x}|\mathbf{y} = C_k)$ denote the class-conditional density of features \mathbf{x} (in this case for class C_k). Consider the data set $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ where as above $\mathbf{y}_i \in \{C_k\}_{k=1}^K$ is encoded as a one-hot target vector and the data are independent.

1. Write out the log-likelihood of the data set, $\ln p(D; \boldsymbol{\pi})$.
2. Since the prior forms a distribution, it has the constraint that $\sum_k \pi_k - 1 = 0$. Using the hint on Lagrange multipliers below, give the expression for the maximum-likelihood estimator for the prior class-membership probabilities, i.e. $\hat{\pi}_k$. Make sure to write out the intermediary equation you need to solve to obtain this estimator. Briefly state why your final answer is intuitive.

For the remaining questions, let the class-conditional probabilities be Gaussian distributions with the same covariance matrix

$$p(\mathbf{x}|\mathbf{y} = C_k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \text{ for } k \in \{1, \dots, K\}$$

and different means $\boldsymbol{\mu}_k$ for each class.

3. Derive the gradient of the log-likelihood with respect to vector $\boldsymbol{\mu}_k$. Write the expression in matrix form as a function of the variables defined throughout this exercise. Simplify as much as possible for full credit.
4. Derive the maximum-likelihood estimator $\hat{\boldsymbol{\mu}}_k$ for vector $\boldsymbol{\mu}_k$. Briefly state why your final answer is intuitive.
5. Derive the gradient for the log-likelihood with respect to the covariance matrix $\boldsymbol{\Sigma}$ (i.e., looking to find an MLE for the covariance). Since you are differentiating with respect to a *matrix*, the resulting expression should be a matrix!
6. Derive the maximum likelihood estimator $\hat{\boldsymbol{\Sigma}}$ of the covariance matrix.

Hint: Lagrange Multipliers. Lagrange Multipliers are a method for optimizing a function f with respect to an equality constraint, i.e.

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } g(\mathbf{x}) = 0.$$

This can be turned into an unconstrained problem by introducing a Lagrange multiplier λ and constructing the Lagrangian function,

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}).$$

It can be shown that it is a necessary condition that the optimum is a critical point of this new function. We can find this point by solving two equations:

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial \mathbf{x}} = 0 \quad \text{and} \quad \frac{\partial L(\mathbf{x}, \lambda)}{\partial \lambda} = 0$$

Cookbook formulas. Here are some formulas you might want to consider using to compute difficult gradients. You can use them in the homework without proof. If you are looking to hone your matrix calculus skills, try to find different ways to prove these formulas yourself (will not be part of the evaluation of this homework). In general, you can use any formula from the matrix cookbook, as long as you cite it. We opt for the following common notation: $\mathbf{X}^{-\top} := (\mathbf{X}^{\top})^{-1}$

$$\begin{aligned} \frac{\partial \mathbf{a}^{\top} \mathbf{X}^{-1} \mathbf{b}}{\partial \mathbf{X}} &= -\mathbf{X}^{-\top} \mathbf{a} \mathbf{b}^{\top} \mathbf{X}^{-\top} \\ \frac{\partial \ln |\det(\mathbf{X})|}{\partial \mathbf{X}} &= \mathbf{X}^{-\top} \end{aligned}$$

Solution

1. Write out the log-likelihood of the data set, $\ln p(D; \boldsymbol{\pi})$.

$$p(D; \boldsymbol{\pi}) = p(\mathbf{x}, \mathbf{y} | \boldsymbol{\pi})$$

$$p(\mathbf{x}, \mathbf{y} | \boldsymbol{\pi}) = p(\mathbf{x} | \mathbf{y}, \boldsymbol{\pi}) p(\mathbf{y} | \boldsymbol{\pi})$$

$$p(\mathbf{x}, \mathbf{y} | \boldsymbol{\pi}) = \prod_{i=1}^n p(\mathbf{x} | \mathbf{y} = y_i, \boldsymbol{\pi}) p(\mathbf{y} = y_i | \boldsymbol{\pi})$$

$$p(\mathbf{x}, \mathbf{y} | \boldsymbol{\pi}) = \prod_{i=1}^n \prod_{k=1}^K [p(x | y = C_k, \boldsymbol{\pi}) \pi_k]^{y_{ik}}$$

$$\ln p(\mathbf{x}, \mathbf{y}; \boldsymbol{\pi}) = \sum_{i=1}^n \sum_{k=1}^K y_{ik} \ln(p(x_n | y_n = C_k)) + y_{ik} \ln(\pi_k)$$

where $y_{ik} = \mathbb{I}(y_i = C_k)$.

2. Since the prior forms a distribution, it has the constraint that $\sum_k \pi_k - 1 = 0$. Using the hint on Lagrange multipliers below, give the expression for the maximum-likelihood estimator for the prior class-membership probabilities, i.e. $\hat{\pi}_k$. Make sure to write out the intermediary equation you need to solve to obtain this estimator. Briefly state why your final answer is intuitive.

Answer:

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ s.t. } g(\mathbf{x}) = 0$$

$$f(\boldsymbol{\pi}) = \sum_{i=1}^n \sum_{k=1}^K y_{ik} \ln(p(x_n | y_n = C_k)) + y_{ik} \ln(\pi_k)$$

$$g(\boldsymbol{\pi}) = \sum_k \pi_k - 1$$

$$L(\boldsymbol{\pi}, \lambda) = f(\boldsymbol{\pi}) + \lambda g(\boldsymbol{\pi})$$

$$\frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \boldsymbol{\pi}} = 0 \quad \text{and} \quad \frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \lambda} = 0$$

$$\frac{\partial L(\boldsymbol{\pi}, \lambda)}{\partial \boldsymbol{\pi}} = \begin{bmatrix} \frac{\partial L}{\partial \pi_1} \\ \vdots \\ \frac{\partial L}{\partial \pi_K} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial \pi_k} = \left[\sum_{i=1}^n \frac{\mathbb{I}_y}{\pi_k} \right] = \vec{0}$$

$$\lambda + \sum_{i=1}^n \frac{\mathbb{I}_y}{\pi_k} = 0$$

$$\lambda = -n$$

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i = C_k)$$

Intuitively, the best-guess estimate for class probabilities is the percentage of observations in each class.

3.

$$\nabla_{\mu_k}(\ln p(D|\boldsymbol{\pi})) = \sum_{i=1}^n \sum_{k=1}^K \nabla_{\mu_k} \ln \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

$$\nabla_{\mu_k}(\ln p(D|\boldsymbol{\pi})) = \sum_{i=1}^n \sum_{k=1}^K \nabla_{\mu_k} (\mathbb{I}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k))$$

$$\frac{d}{d\vec{x}} (\vec{x}^T A \vec{x}) = (A + A^T) \vec{x}$$

$$= \sum_{k=1}^K \mathbb{I}(-\frac{1}{2})(\Sigma^{-1} + \Sigma^{-T})(-1)(x - \mu_k) = \sum_{k=1}^K \mathbb{I}(\Sigma^{-1})(x - \mu_k)$$

4.

$$\nabla_{\mu_k} p(D|\boldsymbol{\pi}) = 0 \Rightarrow \sum_{k=1}^K \mathbb{I}(\Sigma^{-1})(\mathbf{x} - \mu_k) = 0$$

$$\Sigma^{-1} \mathbf{x} = \Sigma^{-1} \mu_k \Rightarrow \Sigma \Sigma^{-1} \mathbf{x} = \Sigma \Sigma^{-1} \mu_k$$

$$\Rightarrow \frac{1}{\sum_{k=1}^K \mathbb{I}_k} \sum_{k=1}^K \mathbb{I}_k \mathbf{x} = \hat{\mu}_k$$

In other words, the best estimate for the vector of mean features for each class is the average of the feature values for all the observations in that class. This makes sense because you would expect further observations from that class to be similar to the ones you've already seen.

5.

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = \sum_{i=1}^n \sum_{k=1}^K \ln \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = \sum_{i=1}^n \sum_{k=1}^K \nabla_{\Sigma} \ln \left(\det(2\pi\Sigma)^{-1/2} e^{-(\mathbf{x}-\mu_k)^T (\Sigma^{-1})(\mathbf{x}-\mu_k)/2} \right)$$

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = \sum_{i=1}^n \sum_{k=1}^K -\frac{1}{2} \nabla_{\Sigma} \ln (\det(2\pi\Sigma)) + \nabla_{\Sigma} \ln \left(e^{-(\mathbf{x}-\mu_k)^T (\Sigma^{-1})(\mathbf{x}-\mu_k)/2} \right)$$

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = -\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \nabla_{\Sigma} \ln(\det(2\pi\Sigma)) + \nabla_{\Sigma} ((\mathbf{x} - \mu_k)^T (\Sigma^{-1})(\mathbf{x} - \mu_k))$$

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = -\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K (\det(2\pi\Sigma))^{-T} + \nabla_{\Sigma} ((\mathbf{x} - \mu_k)^T (\Sigma^{-1})(\mathbf{x} - \mu_k))$$

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = -\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \left((2\pi\Sigma)^{-T} + \Sigma^{-T}(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T \Sigma^{-T} \right)$$

6.

$$\nabla_{\Sigma}(\ln p(D|\boldsymbol{\pi})) = 0$$

$$-2\pi\Sigma^{-T} = \Sigma^{-T}(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T \Sigma^{-T}$$

Right-multiplying both sides of the equation by Σ^T ,

$$-2\pi I = \Sigma^{-T}(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T$$

Left-multiplying by Σ^T ,

$$-2\pi\Sigma^T = (\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T$$

We know that Σ is symmetric because it is a covariance matrix, so:

$$\Sigma = -\frac{1}{2\pi}(\mathbf{x} - \mu_k)(\mathbf{x} - \mu_k)^T$$

This makes sense because we know from statistics that the covariance equals the expectation of the product of the first variable minus its mean with the second variable minus its mean. This formula above is a generalization to more dimensions of data.

Problem 3 (Classifying Stars, 15pts)

You're tasked with classifying three different kinds of stars using their magnitudes and temperatures. See `star.png` for a plot of the data, adapted from http://astrosci.scimuze.com/stellar_data.htm and available as `data/hr.csv`, which you will find in the Github repository.

The CSV file has three columns: type, magnitude, and temperature. The first few lines look like this:

```
Type,Magnitude,Temperature
Dwarf,-5.8,-0.35
Dwarf,-4.1,-0.31
...
```

In this problem, you will code up 4 different classifiers for this task:

- A three-class generalization of logistic regression**, also known as softmax regression, in which you implement gradient descent on the negative log-likelihood. In Question 2 you will explore the effect of using different values for the learning rate η (`self.eta`) and regularization strength λ (`self.lam`). Make sure to include a bias term and to use L2 regularization. See CS181 Textbook's Chapter 3.6 for details on multi-class logistic regression and softmax. For your implementation, use the loss and gradient expressions provided there.
- A generative classifier with Gaussian class-conditional densities with a *shared covariance matrix*** across all classes. Feel free to re-use your Problem 2 results.
- Another generative classifier with Gaussian class-conditional densities, but now with a *separate covariance matrix*** learned for each class. (Note: The staff implementation can switch between the two Gaussian generative classifiers with just a few lines of code.)
- A kNN classifier** in which you classify based on the $k = 1, 3, 5$ nearest neighbors and the following distance function:

$$\text{dist}(\text{star}_1, \text{star}_2) = ((\text{mag}_1 - \text{mag}_2)/3)^2 + (\text{temp}_1 - \text{temp}_2)^2$$

where nearest neighbors are those with the smallest distances from a given point.

Note 1: When there are more than two labels, no label may have the majority of neighbors. Use the label that has the most votes among the neighbors as the choice of label.

Note 2: The grid of points for which you are making predictions should be interpreted as our test space. Thus, it is not necessary to make a test point that happens to be on top of a training point ignore itself when selecting neighbors.

After implementing the above classifiers, complete the following exercises:

- Plot the decision boundaries generated by each classifier for the dataset. Include them in your PDF. Identify the similarities and differences among the classifiers. What explains the differences?
- For logistic regression only, make a plot with "Number of Iterations" on the x-axis and "Negative Log-Likelihood Loss" on the y-axis for several configurations of the hyperparameters η and λ . Specifically, try the values 0.05, 0.01, and 0.001 for each hyperparameter. Limit the number of gradient descent iterations to 200,000. What are your final choices of learning rate (η) and regularization strength (λ), and why are they reasonable? How does altering these hyperparameters affect the ability to converge, the rate of convergence, and the final loss (a qualitative description is sufficient)? You only need to submit one plot for your final choices of hyperparameters.

Note: The *likelihood* of the model is the probability of data given the model—it should not include the regularization term. The *objective* is the combination of the likelihood and the regularizer.

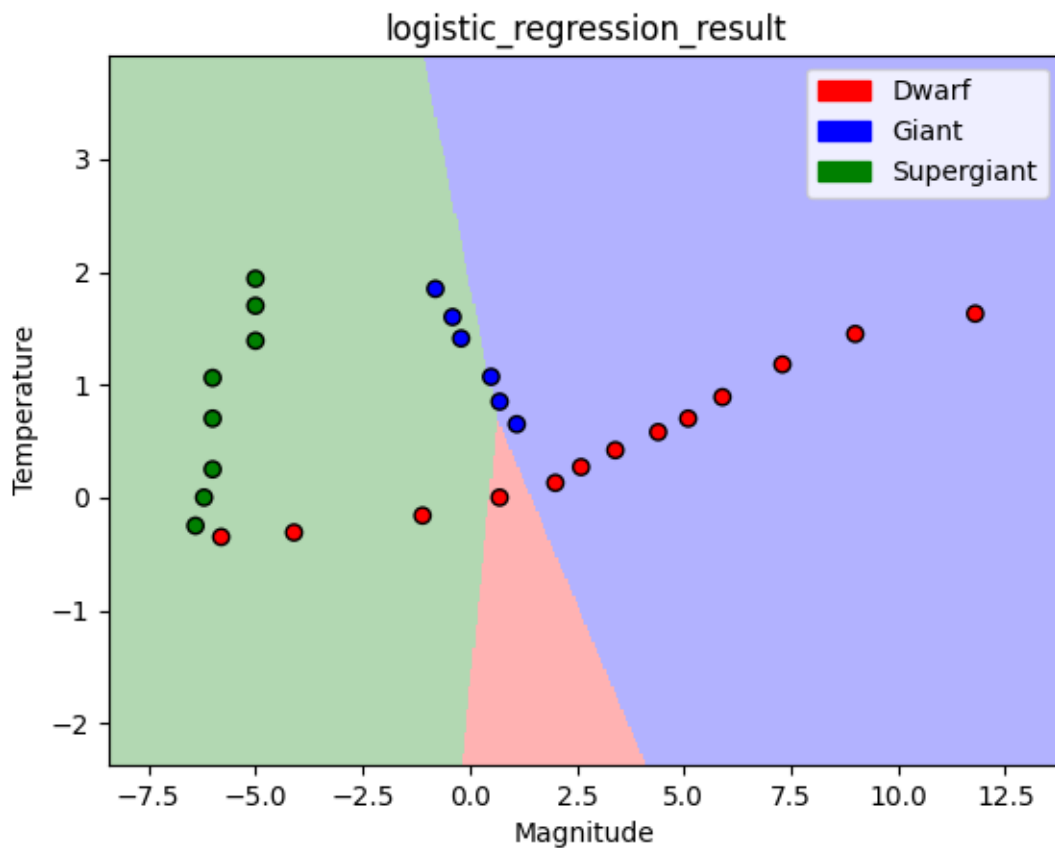
- For both Gaussian generative models, report the negative log-likelihood loss. Which model has a lower loss, and why? For the separate covariance model, be sure to use the covariance matrix that matches the true class of each data point.
- Consider a star with Magnitude 6 and Temperature 2. To what class does each classifier assign this star? Do the classifiers give any indication as to whether or not you should trust them?

Problem 3 (cont.)

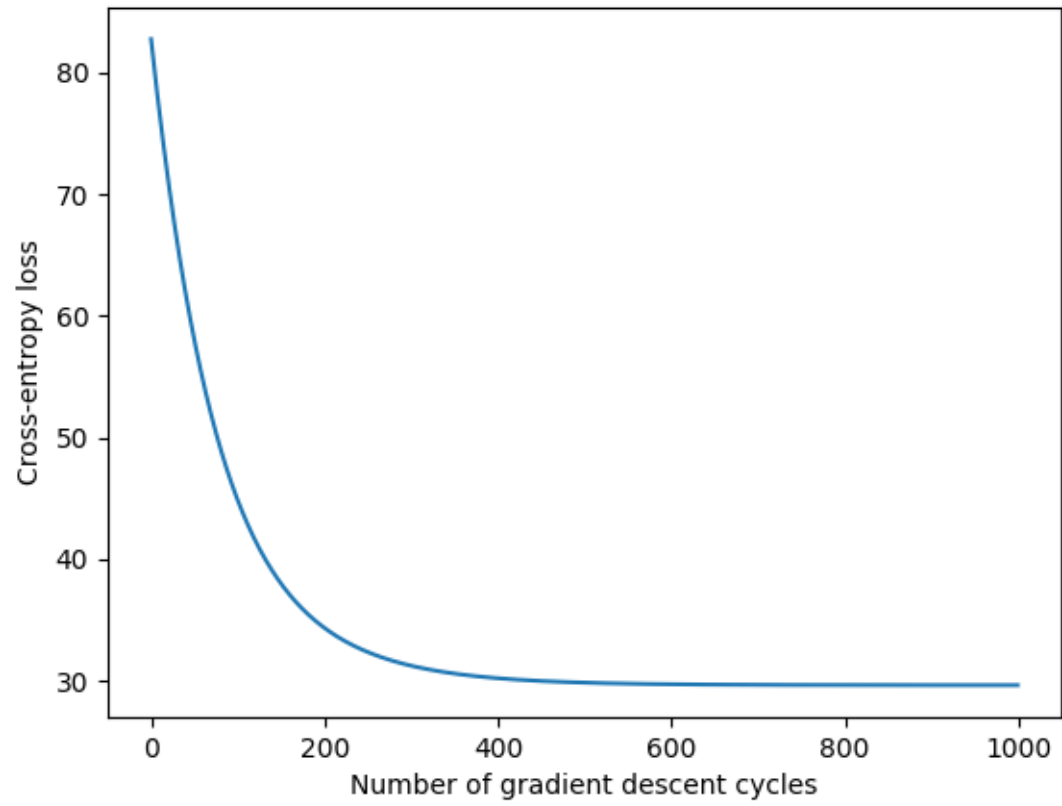
Implementation notes: Run the controller file, `T2_P3.py`, to test your code. Write the actual implementations in the `GaussianGenerativeModel`, `LogisticRegression`, and `KNNModel` classes, which are defined in the three `T2_P3_ModelName.py` files. These classes follow the same interface pattern as `sklearn`. Their code currently outputs nonsense predictions just to show the high-level interface, so you should replace their `predict()` implementations. You'll also need to modify the hyperparameter values in `T2_P3.py` for logistic regression.

Solution

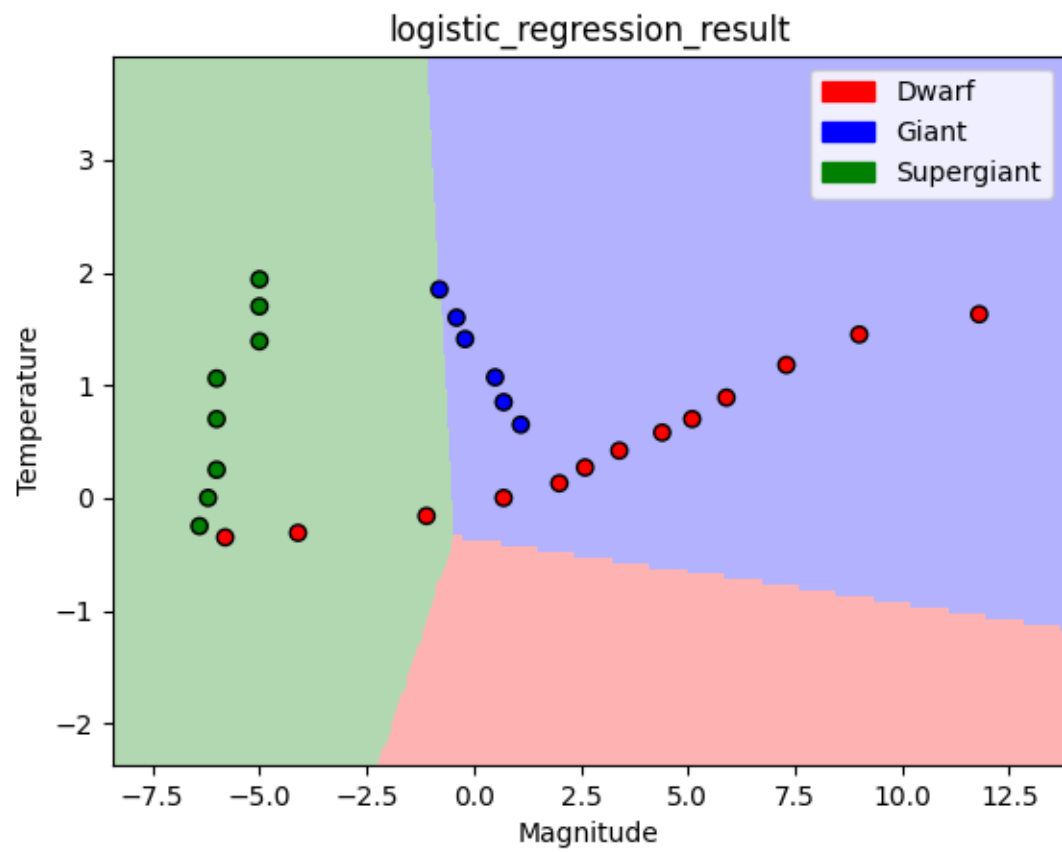
1. 3.1



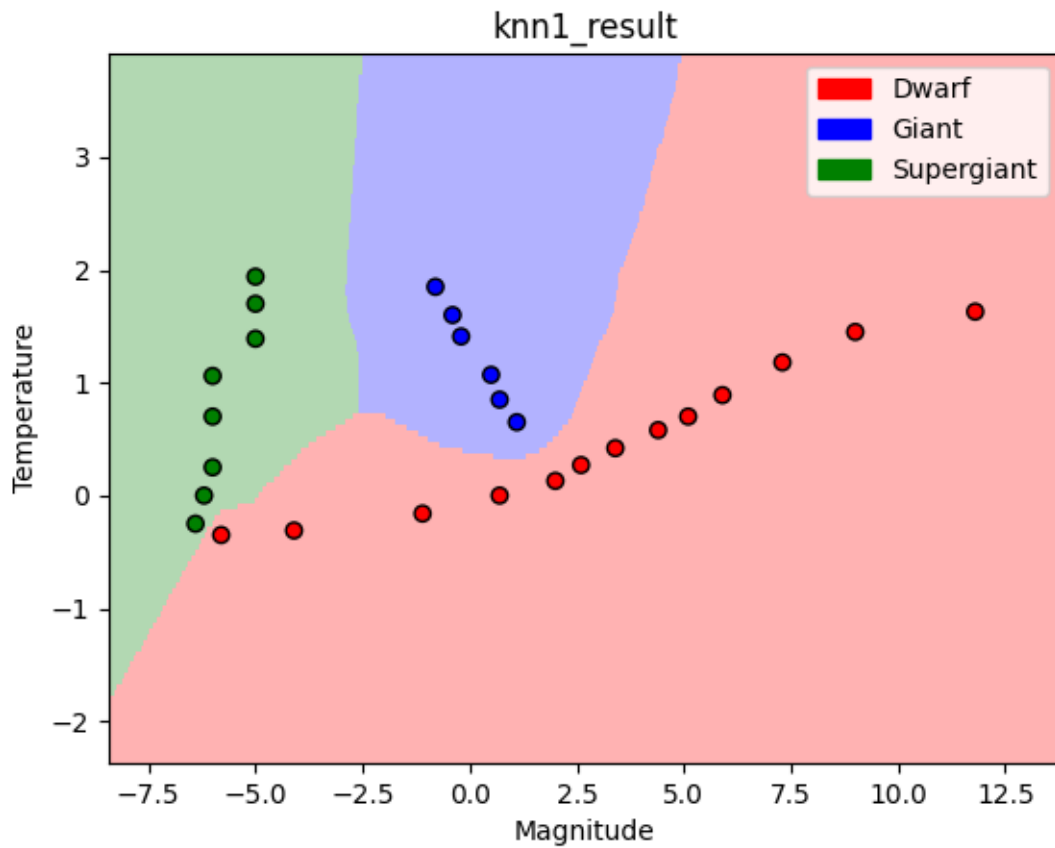
Loss vs number of gradient descent iterations

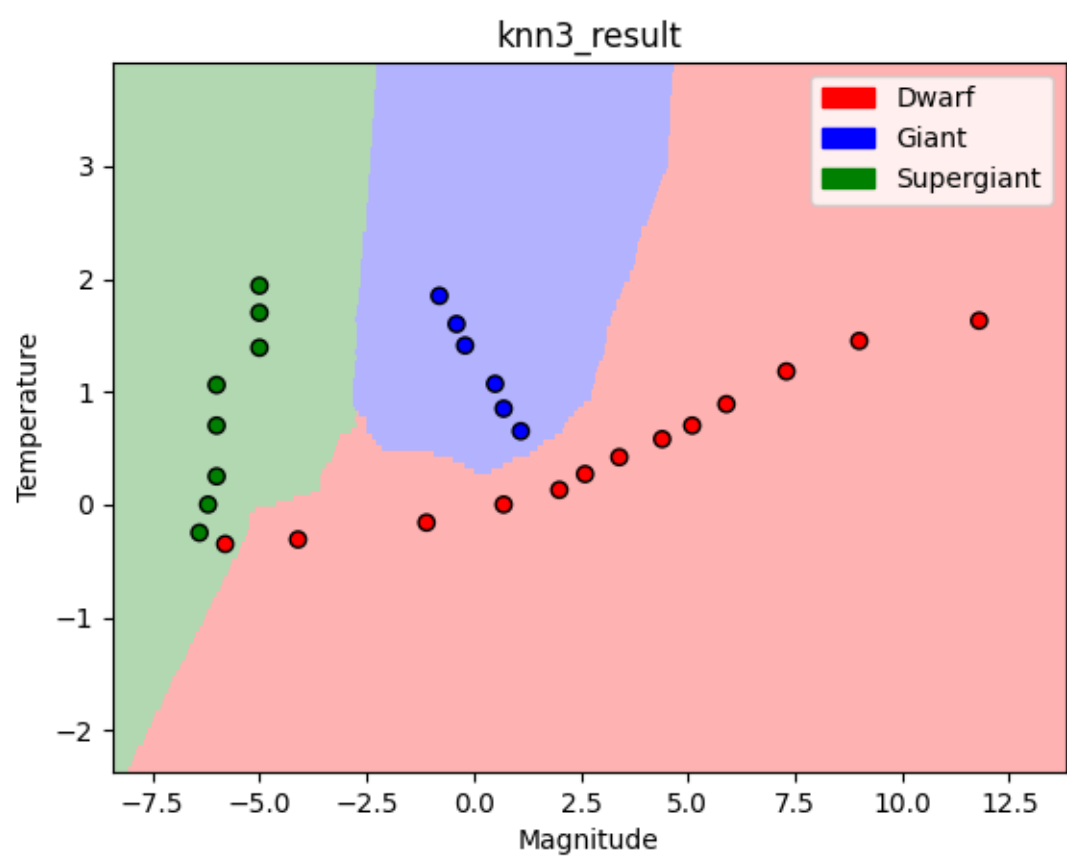


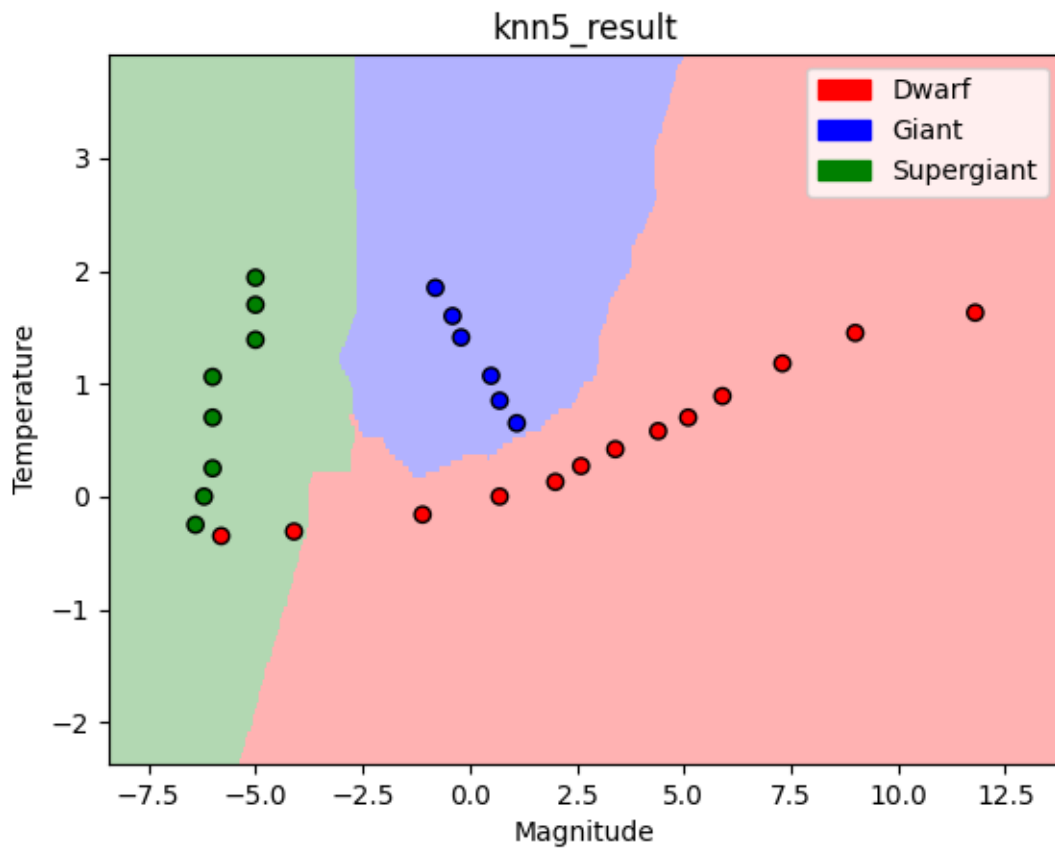
Logistic regression result with regularization penalty:



When the k-NN threshold (number of nearest neighbors) is $k=1$, every point in the boundary plot is within the region of its own color. This is because the model strongly relies on the single nearest points and assumes any new observations match the classes of whichever past training observation was most similar. When k is increased to 3 or 5, more red points can be found in the green region, since they are close to many of the supergiant observations but far from the other dwarf stars. In particular, points can be found in a region categorized as another class if they are closer to k points of that class than they are to $k-1$ other points of their own class.







2. 3.2

3. 3.3

4. 3.4

The classifiers with $k=1$, $k=3$, $k=5$ predict that a star with magnitude 6 and temperature 2 is a dwarf. k -NN classifiers do not give any indication of how confident they are or how they arrived at this result. If classifying in two dimensions, you can at least see how far away a point would have to be (i.e. how big a change in the features) before the classification changed, but this is not generally possible in higher dimensions (without PC decomposition).

Name: Robi Rahman

Collaborators and Resources

Whom did you work with, and did you use any resources beyond cs181-textbook and your notes?

Nikola Jurkovic, Warren Sunada-Wong

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Calibration

Approximately how long did this homework take you to complete (in hours)?

15 hours