DS-GA 3001, Introduction to Data Science (PhD)
Prof. Arthur Spirling, Prof. Julia Kempe
Assignment date: Nov 2, 2020

Homework 2: Solutions

This homework must be returned to your TA by **9pm Eastern Time, November 14, 2020 (Saturday)**. Late work will incur penalties of the equivalent of one third of a letter grade per day late, and **up to three days only**.

It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be in one of the following formats: (1) A set of answers and a clearly commented `Python` code appendix (use comments to identify code relevant to each answer you produced), (2) A report consisting of clearly marked answers, each accompanied by the relevant code (e.g., a report generated using Python markdown or similar). **In either case, your code must be included in full, such that your understanding of the problems can be assessed.**

---

1. Fairness from the point of view of different stakeholders.

   (a) Which of the following are plausible explanations for the False Positive Rate disparity in COMPAS's risk predictions, discussed in this ProPublica article? For each (A)-(D) below, state whether it's a plausible explanation, and briefly justify your answer (1-2 sentences per item): (A) Northpointe failed to make COMPAS blind to race; (B) COMPAS uses machine learning, which picks up and reproduces historical patterns of injustice; (C) COMPAS uses machine learning, which performs worse for minority groups because less training data is available; (D) Northpointe defines recidivism as "a finger-printable arrest involving a charge and a filing of any uniform crime reporting (UCR) code", and it may be the case that members of minority groups are arrested at higher rates.

   (A): **Implausible** if we interpret the question literally, to mean that Northpointe failed to remove "race" from the set of features being considered by COMPAS. We know from the literature that race was not one of the features in the input, yet, there was still an FPR disparity. Further, we also know that simply removing the sensitive feature ("fairness through blindness") rarely helps address disparity, since there are often correlations between features in the data. **Plausible** if we interpret "blinding" to mean using a more sophisticated method to control for correlations between race and other features.

   (B): **Plausible**; indeed, we believe that FPR disparity can be explained well by the fact that historical data reflects disparate treatments of members of different racial groups by the criminal justice system, as reflected in historical data. Even if race was not used directly, a machine learning method will be able to reconstruct historical patterns of injustice through proxies.

(C): **Implausible**. The FPR disparity in question is between the Caucasian and African American populations. We know that there is ample data available for both demographic groups, and in particular that African Americans are represented more prominently in the data than Caucasians.

(D): **Plausible**. Higher arrest rates for African Americans are recorded in the training data, over-estimating recidivism for this demographic group. Thus, the software learns to overestimate recidivism in this group, leading to an FPR disparity.

(b) Consider again the COMPAS investigation by ProPublica, and additionally consider Northpointe's response to ProPublica (you may wish to consult Northpointe's report). For each criterion (A)-(E) below, explain in 1-2 sentences which stakeholders would benefit from a model that optimizes that metric and why. If you believe that a criterion is not reasonable in this case, state so. (A) Accuracy; (B) Positive predictive value; (C) False positive rate; (D) False negative rate; (E) Statistical parity (demographic parity among the individuals receiving any prediction).

(A): Accuracy benefits multiple stakeholders, including the software vendor, the decision maker, and the general public. In a sense, high accuracy, like calibration, is a necessary condition that a risk instrument should meet to be considered useful.

(B): Positive Predictive Value (PPV), defined as TP/(TP+FP), is higher when the true positives constitute a high proportion of the positive class. PPV benefits society, in the sense that we don't spend resources on incarcerating individuals who do not go on to reoffend (FP). It also clearly benefits the prisoners who would not go on to reoffend themselves.

(C): False Positive Rate (FPR), defined as FP/N, is lower when fewer of the defendants who would not go on to reoffend are classified as high risk. Minimizing the FPR benefits low-risk defendants. However, a decision maker may be willing to incur a higher FPR to help lower the false negative rate (FNR)

(D): False Neagtive Rate (FNR), defined as FN/P, is lower when fewer of the defendants who would go on to reoffend are classified as low risk. A low FNR benefits society by ensuring that individuals who are likely to recidivate do not have a chance to do so. It also benefits a decision maker whose reputation will suffer if an individual whom they release goes on to commit a crime. Finally, it benefits the software vendor that caters to this type of a decision maker.

(E): Statistical parity (demographic parity among the individuals receiving any prediction), does not make sense in this application. Note that we can "optimize" statistical parity at many levels of accuracy; statistical parity alone is not unambiguously better for any group. (A system that predicts high risk for every single individual has perfect statistical parity.) Furthermore, we are not interested in equalizing outcomes between

populations per se. Rather, fairness here amounts to balancing the kinds of error that different demographic groups incur—a goal that, as we know from the literature, cannot be met directly and so requires trade-offs.

2. This exercise relates to the bias-variance trade-off derivation we had in Lab7. Assume the standard statistical learning setup: $\mathcal{X}, \mathcal{Y}, \mathcal{A}$ are input, output and action spaces, and $(X, Y) \sim P$ is the joint data distribution; for this exercise, consider the square loss.

   (a) Define a hypothesis space $\mathcal{F} = \mathcal{A}$ of constant functions. Prove that the Bayes decision function form this space is $a^* = \mathbb{E}Y$. Compute the associated Bayes risk.
   The risk associated with constant action $a \in \mathcal{A}$ is $\mathbb{E}(a - Y)^2 = \mathbb{E}(a^2 - 2aY + Y^2)$ $= a^2 - 2a\mathbb{E}Y + \mathbb{E}Y^2 = a^2 - 2a\mathbb{E}Y + \mathrm{Var}(Y) + \mathbb{E}^2 Y = (a - \mathbb{E}Y)^2 + \mathrm{Var}(Y)$. Hence, the minimum of this function is attained when $a = \mathbb{E}Y$, so the Bayes action is $a^* = \mathbb{E}Y$ and the associated risk is $\mathrm{Var}(Y)$.

   (b) Prove that the Bayes decision function (not restricted to $\mathcal{F} = \mathcal{A}$) is $\bar{y}(X) = \mathbb{E}_{Y|X} Y$ (hint: recall from Lab7 that $\mathbb{E}_{(V,W)} g(V, W) = \mathbb{E}_V \mathbb{E}_{W|V} g(V, W)$ for any random variables V, W and any function g).
   Note that, for any function $f \colon \mathcal{X} \to \mathcal{A}$ and any $x \in \mathcal{X}$, part (a) dictates

$$\mathbb{E}_{Y|X=x}(\bar{y}(x) - Y)^2 = \mathrm{Var}(Y) \leq \mathbb{E}_{Y|X=x}(f(x) - Y)^2$$

   since $\bar{y}(x) = \mathbb{E}_{Y|X=x} Y$ is the Bayes decision function for the conditional random variable $Y|X = x$. Now, this implies

$$\begin{aligned}\mathcal{R}(\bar{y}) &= \mathbb{E}_P(\bar{y}(X) - Y)^2 = \mathbb{E}_X \mathbb{E}_{Y|X=x}(\bar{y}(x) - Y)^2 \\ &\leq \mathbb{E}_X \mathbb{E}_{Y|X=x}(f(x) - Y)^2 = \mathbb{E}_P(f(x) - Y)^2 = \mathcal{R}(f),\end{aligned}$$

   meaning that $\bar{y}$ is indeed the Bayes decision function.

3. This exercise refers to the `question3_student.ipynb` Jupyter notebook; you will prepare data for the modeling problem in the next question. The five files distributed together with this assignment are `rpl17.html`, `rpl18.html`, `rpl19.html`, `rpl20.html`, `rpl21.html`, which contain detailed statistics on soccer matches in the last five seasons of the Russian Premiere League. Your task is to complete the `process_season` function. Please follow the more detailed instructions in the notebook.
   See `question3_solution.ipynb`

4. This exercise refers to the `question4_student.ipynb` notebook. You will build, train, optimize and analyze a binary logistic regression model to predict if a soccer match will end in a home win or not using data produced in Question 3 (or you can use `.npy` files provided with this assignment). The binarization of target values is done for you in the beginning of the notebook (positive class corresponds to a home win). Throughout this question, you should not use the `scikit-learn` library.

   (a) The training data is likely unbalanced; upsample the smaller class to achieve a perfectly balanced training dataset (your solution should work both when the smaller class is positive and when it is negative). Since our dataset is still rather small, it is wise to use cross-validation; prepare five different training/validation splits for a 5-fold cross-validation.

(b) Update the `LogisticRegression` class from Homework 1 to allow for $L2$ regularization (note the argument $C = \lambda^{-1}$, where λ is the regularization strength). Recall that the loss function for this model is

$$l(\theta) = -\sum_{i=1}^{n} x_{ij} \left(y_i - \sigma \left(\theta^T \bar{x}_i^T \right) \right) + \lambda \|\theta\|_2^2.$$

(c) Complete the `validate` function that plots the ROC curve for each of the $K = 5$ classifiers (one per fold). In addition, the plot should contain a pointwise average of those ROC curves and a dashed diagonal from $(0, 0)$ to $(1, 1)$. Finally, compute the mean AUC score. Note that you cannot use `scikit-learn` in this question.

(d) Fit Logistic Regression models using 5-fold cross-validation for each value of C provided in the notebook. Choose reasonable values for the learning rate α and `max_iter` (recall that logistic loss is convex, so gradient descent is guaranteed to converge given sufficiently small α; thus, increasing iterations and reducing α is a safe bet). Apply the `validate` function you prepared in the previous question for each selection of C. Display the 8 resulting plots and comment on the optimal choice of C. Note that subplot instances are already are prepared in the code.

(e) Complete the `confusion_matrix`, `precision`, `recall`, and `f1_score` functions. Then, use cross-validation to select the appropriate threshold for prediction by optimizing the mean (across folds) $F1$ score. Use C selected in the previous part. Report your findings.

(f) CSKA will be playing against Rostov in Moscow (home stadium for CSKA) on November 9th. The odds from bookmakers are 1.81 for CSKA and 1.99 for Rostov or a draw (which means that betting \$1 on CSKA could win you \$1.81 (including your bet of \$1), while betting on Rostov or a draw—\$1.99). The current statistics on both teams are prepared for you in the file `CSKA-Rostov.npy`. Train your best model on all training data, compute the confusion matrix on the test set and assess the expected profit brought by your model under these odds (recall that the positive class, in this case, is the victory of CSKA). Then, compute the prediction and check after Novemeber 9th if it was correct!

(g) What are some disadvantages of predicting soccer matches in this way and how the modeling pipeline could be improved?
See `question4_solution.ipynb`

5. In this problem you will relate Logistic Regression (LR) and Support Vector Machines (SVM) for binary classifiers. We will go from LR to SVM. Recall that the LR computes a probability $p_{LR}(y = 1|x) = \sigma(\omega x + b))$ for each data point x. Let us now assume that instead of maximizing MLE to get to the parameter ω, we want to define a classifier from p_{LR} as follows:

$$\hat{y} = \begin{cases} 1 & \text{if} \quad p_{LR}(y = 1|x) \geq p_{LR}(y = 0|x) \\ 0 & \text{otherwise} \end{cases}$$

Now, impose that this classifier, on training data, needs to make the right decision with some *margin of certainty*: either $p_{LR}(y = 1|x) \geq c \cdot p_{LR}(y = 0|x)$ or $p_{LR}(y = 0|x) \geq c \cdot p_{LR}(y = 1|x)$. Among all the solutions for ω we will chose the one that minimizes its L2-norm. Derive that the classifier is a *hard-margin SVM* and give a choice of c to match to the definition of SVM.

Let $c = e$; we will show that the above mathematical optimization problem has the same formulation as that of hard-margin SVM. Indeed, the objective function to minimize here is $\|w\|_2$ subject to $p_{LR}(y_i = 1|x_i) \geq cp_{LR}(y_i = 0|x_i)$ if $y_i = 1$ and $cp_{LR}(y_i = 1|x_i) \leq p_{LR}(y_i = 0|x_i)$ if $y_i = 0$ for all $i \in [n]$. Note that $p_{LR}(y_i = 1|x_i) = \sigma(x\omega + b) = 1 - p_{LR}(y_i = 0|x_i)$, so the above constraints can be rewritten as

$$\begin{cases} \frac{1}{1+e^{-x_i\omega-b}} \geq c\left(1 - \frac{1}{1+e^{-x_i\omega-b}}\right) = \frac{ce^{-x_i\omega-b}}{1+e^{-x_i\omega-b}} & \text{if } y_i = 1, \\ \frac{c}{1+e^{-x_i\omega-b}} \leq \left(1 - \frac{1}{1+e^{-x_i\omega-b}}\right) = \frac{e^{-x_i\omega-b}}{1+e^{-x_i\omega-b}} & \text{if } y_i = 0, \end{cases}$$

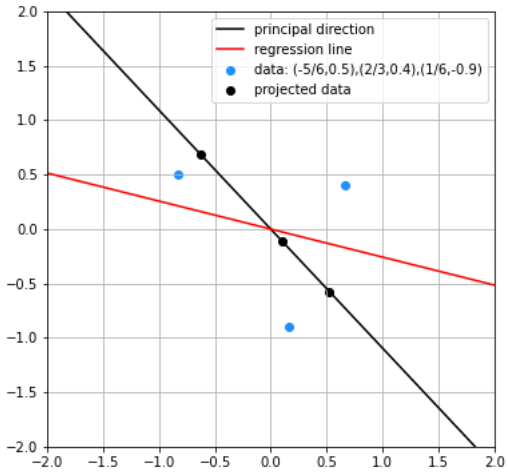
which (plugging in $c = e$) is equivalent to

$$\begin{cases} ce^{-x_i\omega-b} \leq 1 & \text{if } y_i = 1, \\ c \leq e^{-x_i\omega-b} & \text{if } y_i = 0 \end{cases} \quad \begin{cases} e \leq e^{x_i\omega+b} & \text{if } y_i = 1, \\ e \leq e^{-x_i\omega-b} & \text{if } y_i = 0 \end{cases} \quad \begin{cases} 1 \leq x_i\omega + b & \text{if } y_i = 1, \\ 1 \leq -x_i\omega - b & \text{if } y_i = 0, \end{cases}$$

which is the same as $1 \leq (x_i\omega + b)(2y_i - 1)$ for all $i \in [n]$ (transforming labels by $1 \mapsto 1$ and $0 \mapsto -1$), which are exactly the hard-margin SVM constraints, while the optimization function has the same global optimum.

6. SVD - Lily's question:

   (a) Explicitly show with a small example (e.g. 3-4 points) that the least squares regression line differs from the first PCA direction.



   (b) Further, in 2D, show that PCA minimizes *perpendicular L2-distance* between the first component line and the data points.
   Note that the first principal vector specifies the direction of the projection in the original feature space, thus corresponding to a line passing through the origin. This line will not minimize the sum of squared distances from the points in general unless an appropriate intercept is chosen. In fact, it can be shown that this line should pass through (μ_X, μ_Y). Alternatively, without loss of generality, we could just as well center the data

(i.e. subtract (μ_X, μ_Y)) and prove that $ax + by = 0$ minimizes the sum of squared distances where $\bar{w} = (a, b)$ is the first principal direction. Thus, assume the data is centered: $\frac{1}{n}\sum_{i=1}^n \bar{x}_i = 0$. The first principal direction is a solution to

$$\max_{\|\bar{w}\|=1} \text{Var}\left(\text{proj}_{\bar{w}}\bar{x}_i\right) = \max_{\|\bar{w}\|=1} \text{Var}\langle \bar{x}_i, \bar{w}\rangle = \max_{\|\bar{w}\|=1} \left[\frac{1}{n}\sum_{i=1}^n \langle \bar{x}_i, \bar{w}\rangle^2 - \left(\frac{1}{n}\sum_{i=1}^n \langle \bar{x}_i, \bar{w}\rangle\right)^2\right]$$
$$= \max_{\|\bar{w}\|=1} \frac{1}{n}\sum_{i=1}^n \langle \bar{x}_i, \bar{w}\rangle^2$$

because the last term in the first line is zero as the data is centered. We will now show that the solution to the above problem is also a solution to the minimization of the sum of squared distances. Note that the distance from $\bar{x}_i$ to the line given by $\bar{w}$ is $\|\bar{x}_i - \text{proj}_{\bar{w}}\bar{x}_i\| = \|\bar{x}_i - \langle \bar{x}_i, \bar{w}\rangle \bar{w}\|$ assuming that $\|\bar{w}\| = 1$. Then, the minimization problem is

$$\min_{\|\bar{w}\|=1} \frac{1}{n}\sum_{i=1}^n \|\bar{x}_i - \langle \bar{x}_i, \bar{w}\rangle \bar{w}\|^2 = \min_{\|\bar{w}\|=1} \frac{1}{n}\sum_{i=1}^n \langle \bar{x}_i - \langle \bar{x}_i, \bar{w}\rangle \bar{w}, \bar{x}_i - \langle \bar{x}_i, \bar{w}\rangle \bar{w}\rangle$$
$$= \min_{\|\bar{w}\|=1} \frac{1}{n}\sum_{i=1}^n \|\bar{x}_i\|^2 - 2\langle \bar{w}, \bar{x}_i\rangle^2 + \langle \bar{x}_i, \bar{w}\rangle \|\bar{w}\|^2$$
$$= \max_{\|\bar{w}\|=1} \frac{2}{n}\sum_{i=1}^n \langle \bar{x}_i, \bar{w}\rangle^2,$$

because $\frac{1}{n}\sum_{i=1}^n \|\bar{x}_i\|^2$ is constant, while $\frac{1}{n}\sum_{i=1}^n \langle \bar{x}_i, \bar{w}\rangle \|\bar{w}\|^2 = \frac{1}{n}\sum_{i=1}^n \langle \bar{x}_i, \bar{w}\rangle = \langle \bar{w}, \frac{1}{n}\sum_{i=1}^n \bar{x}_i\rangle = 0$ because data is centered. It is now clear that the two optimization problems have the same solution set, meaning that the line given by the direction of the first principal component is the one minimizing the sum of squared distances from points.

7. PCA: Assume your data set consists of M points of dimension p coming from the following distribution: (1) an index $i \in \{1, \ldots, p\}$ is picked uniformly at random. Then (2) the data point is the vector $a\delta_i = (0, \ldots, 0, a, 0, \ldots, 0)$, where a is in the ith position and comes from an arbitrary distribution $P(a)$ (let's say it is bounded). Compute the last principal vector of the design matrix and show that all other eigenvalues are equal (Hint: the covariance matrix is of the form $C_{i,j} = \lambda + \mu\delta_{ij}$ where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise). Discuss whether PCA is a good way to select features for this problem.

Assume P is discrete with mean $\mu \neq 0$, variance σ^2 and support $\mathcal{X}$. Let $Z \sim \text{Unif}([p])$ be a discrete uniform random variable. Then, $\mathbb{E}X_i = \sum_{x\in\mathcal{X}} x\mathbb{P}(X_i = x) = \sum_{x\in\mathcal{X}} x\mathbb{P}(Z = i)P(x)$ $= p^{-1}\sum_{x\in\mathcal{X}} xP(x) = p^{-1}\mu$ and $\mathbb{E}X_i^2 = \sum_{x\in\mathcal{X}} x^2\mathbb{P}(X_i = x) = \sum_{x\in\mathcal{X}} x^2\mathbb{P}(Z = i)P(x)$ $= p^{-1}\sum_{x\in\mathcal{X}} x^2P(x) = p^{-1}(\sigma^2 + \mu^2)$, so $\text{Var}(X_i) = \mathbb{E}X_i^2 - \mathbb{E}^2X_i = p^{-1}(\sigma^2 + \mu^2) - p^{-2}\mu^2$. Therefore, the ij-th entry of the covariance matrix of X is $\mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j = 0 - \mathbb{E}^2X_i = -p^{-2}\mu^2$ when $i \neq j$ and $\text{Var}(X_i) = p^{-1}(\sigma^2 + \mu^2) - p^{-2}\mu^2$ otherwise. Let $a = p^{-1}(\sigma^2 + \mu^2) - p^{-2}\mu^2$ be the diagonal entry and $b = -p^{-2}\mu^2$ be the off-diagonal entry of the covariance matrix Σ. Then, it is clear that $\bar{v} = \bar{1}$ is an eigenvector with a corresponding eigenvalue $a + (p-1)b$ because $\Sigma\bar{v} = \Sigma\bar{1}$ is a vector with $a + (p-1)b$ in all components (sum of each row of Σ). Further, note that $\bar{v} \in \mathbb{R}^p$ where $\langle \bar{v}, \bar{1}\rangle = 0$ (i.e. the sum of all components of $\bar{v}$ is 0) is an

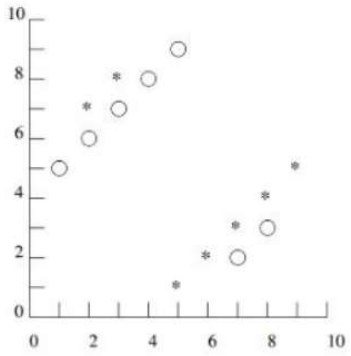
eigenvector corresponding to the eigenvalue of $a - b$. Indeed, for any $k \in [p]$, $(\Sigma\bar{v})_k = \langle \Sigma_k, v \rangle = b\left(\sum_{i=1}^{p} v_i\right) - bv_k + av_k = (a-b)v_k$. Note that there are $p-1$ linearly independent vectors $\bar{v}$ of this form, all of which are eigenvectors corresponding to $a - b$. Moreover, $a + (p-1)b = a + pb - b < a - b$ because $b = -p^{-2}\mu^2 < 0$. Hence, we conclude that there are $p-1$ equal eigenvalues and one smaller eigenvalue. The last principal direction PC_p is given by the eigenvector corresponding to the smallest eigenvalue, so $PC_p = \bar{1}\|\bar{1}\|^{-1} = (p^{-1/2}, p^{-1/2}, \ldots, p^{-1/2})$

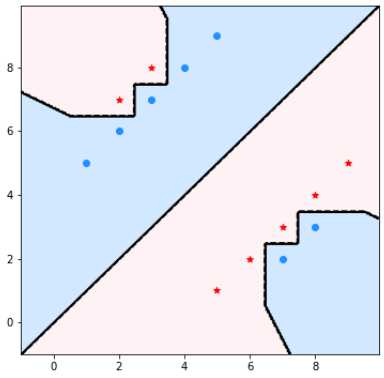
8. k-NN:

   (a) Let k-NN(S) be the k Nearest Neighbor classification algorithm on sample set S, which takes the majority of the closest k points where there are 2 classes (positive and negative).

   - Show that if in both 1-NN(S1) and 1-NN(S2) the label of point x is positive, then in 1-NN($S1 \cup S2$) the label of x is positive.
     Let y_1, y_2 be nearest points from $S1, S2$ to x, respectively. Since 1-NN(S1) and 1-NN(S2) classify x as positive, y_1 and y_2 must have positive labels. Finally, note that either y_1 or y_2 will be the nearest point from $S1 \cup S2$ to x. Thus, 1-NN($S1 \cup S2$) must classify x as positive.
   - Show an example such that in both 3-NN(S1) and 3-NN(S2) the label of x is positive, and in 3-NN($S1 \cup S2$) the label of x is negative.
     $S1 = \{(1,-), (5,+), (6,+)\}$, $S2 = \{(2,-), (3,+), (4,+)\}$ and $x = 0$.

   (b) One of the problems with k-nearest neighbor learning is selecting a value for k. Say you are given the following data set. This is a binary classification task in which the instances are described by two real-valued attributes.





   - What value of k minimizes training set error for this data set, and what is the resulting training set error? Why is training set error not a reasonable estimate of test set error, especially given this value of k?
     With $k = 1$, all training points are classified based on their own label, yielding no mistakes (zero error). When $k = 1$, the algorithm overfits to an extreme extent, so training error is clearly not a good estimate of generalizability.
   - What value of k minimizes the leave-one-out cross-validation error for this data set, and what is the resulting error? Why is cross-validation a better measure of test

set performance? (Note: leave-one-out cross validation on n points trains on $n-1$ points and tests on the last point, for each way to remove one point. It hence runs the model n times.)
mean (across folds) validation error rates for different odd values of k: (1) 10/14, (3): 6/14, (5): 4/14, (7): 4/14, (9): 14/14, (11): 14/14, (13): 14/14. Hence, the lowest error is associated with $k = 5, 7$.

- Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?
  Too large values of k will force prediction to consider a larger neighborhood around an input point, increasing chances that points from the *other/wrong side* will affect the prediction. Too small values of k ($k = 1, 3$) will result in overfitting, i.e., paying attention to those pairs of *adverse* points on each side while making predictions for a point from an otherwise globally dominant class.
- Sketch the 1-nearest neighbor decision boundary for this dataset.
  See above.