

CPSC 340 Assignment 4 (Due Sunday 2021-11-07 at 11:59pm)

Important: Submission Format [5 points]

Please make sure to follow the submission instructions posted on the course website. We will deduct marks if the submission format is incorrect, or if you're not using L^AT_EX and your handwriting is *at all* difficult to read – at least these 5 points, more for egregious issues. Compared to assignment 1, your name and student number are no longer necessary (though it's not a bad idea to include them just in case, especially if you're doing the assignment with a partner).

1 Convex Functions [15 points]

Recall that convex loss functions are typically easier to minimize than non-convex functions, so it's important to be able to identify whether a function is convex.

Show that the following functions are convex:

1. $f(w) = \alpha w^2 - \beta w + \gamma$ with $w \in \mathbb{R}, \alpha \geq 0, \beta \in \mathbb{R}, \gamma \in \mathbb{R}$ (1D quadratic).

Answer: $f'(w) = 2\alpha w - \beta$

$f''(w) = 2\alpha$

Since $\alpha \geq 0$ as specified in the question, this means that $f''(w) \geq 0$, meaning that $f(w)$ is convex, as desired.

2. $f(w) = -\log(\alpha w)$ with $\alpha > 0$ and $w > 0$ (“negative logarithm”)

Answer: $f'(w) = -\frac{1}{w}$

$f''(w) = \frac{1}{w^2}$

Since $w^2 \geq 0$, this means that $f''(w) \geq 0$, meaning that $f(w)$ is convex, as desired.

3. $f(w) = \|Xw - y\|_1 + \frac{\lambda}{2}\|w\|_1$ with $w \in \mathbb{R}^d, \lambda \geq 0$ (L1-regularized robust regression).

Answer: Norms are always convex (as per slide 37 from the L13 slides), and the sum of convex functions is also convex (as per slide 38 from the L13 slides). Since $f(w)$ is a sum of norms, $f(w)$ must be convex, as desired.

4. $f(w) = \sum_{i=1}^n \log(1 + \exp(-y_i w^T x_i))$ with $w \in \mathbb{R}^d$ (logistic regression).

Answer: $f'(w) = \frac{-y_i x_i \exp(-y_i w^T x_i)}{1 + \exp(-y_i w^T x_i)} = \frac{-y_i x_i \exp(-y_i w^T x_i) \exp(y_i w^T x_i)}{1 + \exp(y_i w^T x_i)}$ (from the hint)
 $= \frac{-y_i x_i}{1 + \exp(y_i w^T x_i)}$

$f''(w) = \frac{y_i x_i (y_i x_i \exp(y_i w^T x_i))}{(1 + \exp(y_i w^T x_i))^2} = \frac{(y_i x_i)^2 \exp(y_i w^T x_i)}{(1 + \exp(y_i w^T x_i))^2}$.

Since any squared value is a positive number, and $\exp(x) > 0$ where x is any real number, $f''(w)$ is greater than 0, meaning that $\log(1 + \exp(-y_i w^T x_i))$ is convex. The function $f(w)$ is a summation of these terms, and the summation of convex terms is also convex, so $f(w)$ must also be convex, as desired.

5. $f(w) = \sum_{i=1}^n [\max\{0, |w^T x_i - y_i|\} - \epsilon] + \frac{\lambda}{2}\|w\|_2^2$ with $w \in \mathbb{R}^d, \epsilon \geq 0, \lambda \geq 0$ (support vector regression).

Answer: The expression $|w^T x_i - y_i|$ is the absolute value of a linear function, which is convex. We are taking the maximum between this expression and 0 which is also convex, so the resulting expression must also be convex. We are then taking a summation of these expressions which is the summation of convex expressions, so the result should also be convex. Finally, we are adding this to a squared L-2 norm (convex) which is being multiplied by a non-negative scalar (still convex), which also gives a convex expression. Given all of this information, $f(w)$ must also be convex, as desired.

General hint: for the first two you can check that the second derivative is non-negative since they are one-dimensional. For the last 3, it's easier to use some of the results regarding how combining convex functions can yield convex functions; which can be found in the lecture slides.

Hint for part 4 (logistic regression): this function may at first seem non-convex since it contains $\log(z)$ and \log is concave, but note that $\log(\exp(z)) = z$ is convex despite containing a log. To show convexity, you can reduce the problem to showing that $\log(1 + \exp(z))$ is convex, which can be done by computing the second derivative. It may simplify matters to note that $\frac{\exp(z)}{1+\exp(z)} = \frac{1}{1+\exp(-z)}$.

2 Logistic Regression with Sparse Regularization [30 points]

If you run `python main.py -q 2`, it will:

1. Load a binary classification dataset containing a training and a validation set.
2. Standardize the columns of X , and add a bias variable (in `utils.load_dataset`).
3. Apply the same transformation to `Xvalidate` (in `utils.load_dataset`).
4. Fit a logistic regression model.
5. Report the number of features selected by the model (number of non-zero regression weights).
6. Report the error on the validation set.

Logistic regression does reasonably well on this dataset, but it uses all the features (even though only the prime-numbered features are relevant) and the validation error is above the minimum achievable for this model (which is 1 percent, if you have enough data and know which features are relevant). In this question, you will modify this demo to use different forms of regularization to improve on these aspects.

Note: your results may vary slightly, depending on your software versions, the exact order you do floating-point operations in, and so on.

2.1 L2-Regularization [5 points]

In `linear_models.py`, you will find a class named `LogRegClassifier` that defines the fitting and prediction behaviour of a logistic regression classifier. As with ordinary least squares linear regression, the particular choice of a function object (`fun_obj`) and an optimizer (`optimizer`) will determine the properties of your output model. Your task is to implement a logistic regression classifier that uses L2-regularization on its weights. Go to `fun_obj.py` and complete the `LogisticRegressionLossL2` class. This class' constructor takes an input parameter λ , the L2 regularization weight. Specifically, while `LogisticRegressionLoss` computes

$$f(w) = \sum_{i=1}^n \log(1 + \exp(-y_i w^T x_i)),$$

your new class `LogisticRegressionLossL2` should compute

$$f(w) = \sum_{i=1}^n [\log(1 + \exp(-y_i w^T x_i))] + \frac{\lambda}{2} \|w\|^2.$$

and its gradient. Submit your function object code. Using this new code with $\lambda = 1$, report how the following quantities change: (1) the training (classification) error, (2) the validation (classification) error, (3) the number of features used, and (4) the number of gradient descent iterations.

```
1  class LogisticRegressionLossL2(LogisticRegressionLoss):  
2      def __init__(self, lammy):  
3          super().__init__()  
4          self.lammy = lammy  
5  
6      def evaluate(self, w, X, y):  
7          w = ensure_1d(w)  
8          y = ensure_1d(y)  
9  
10     """YOUR CODE HERE FOR Q2.1"""  
11     #raise NotImplementedError()  
12     Xw = X @ w
```

```

13     yXw = y * Xw
14
15     # Calculate the function value
16     f = np.sum(np.log(1 + np.exp(-yXw))) + (self.lammy/2)*(w @ w)
17
18     # Calculate the gradient value
19     s = -y / (1 + np.exp(yXw))
20     g = X.T @ s + self.lammy*w
21
22     return f, g

```

Answer:

- (1) The training error increased from 0 to 0.002
- (2) The validation error decreased from 0.082 to 0.074
- (3) The number of features used remained 101
- (4) The number of gradient decent iterations decreased from 89 to 30

Note: as you may have noticed, `lambda` is a special keyword in Python, so we can't use it as a variable name. Some alternative options: `lammy` (what Mike's niece calls her toy stuffed lamb), `lamda`, `reg_wt`, λ if you feel like typing it, the sheep emoji¹,

2.2 L1-Regularization and Regularization Path [5 points]

L1-regularized logistic regression classifier has the following objective function:

$$f(w) = \sum_{i=1}^n [\log(1 + \exp(-y_i w^T x_i))] + \lambda \|w\|_1.$$

Because the L1 norm isn't differentiable when any elements of w are 0 – and that's *exactly what we want to get* – standard gradient descent isn't going to work well on this objective. There is, though, a similar approach called *proximal gradient descent* that does work here.

This is implemented for you in the `GradientDescentLineSearchProxL1` class inside `optimizers.py`. Note that to use it, you *don't include the L1 penalty in your loss function object*; the optimizer handles that itself.

Write and submit code to instantiate `LogRegClassifier` with the correct function object and optimizer for L1-regularization. Using this linear model, obtain solutions for L1-regularized logistic regression with $\lambda = 0.01$, $\lambda = 0.1$, $\lambda = 1$, $\lambda = 10$. Report the following quantities per each value of λ : (1) the training error, (2) the validation error, (3) the number of features used, and (4) the number of gradient descent iterations.

```

1 def q2_2():
2     data = utils.load_dataset("logisticData")
3     X, y = data["X"], data["y"]
4     X_valid, y_valid = data["Xvalid"], data["yvalid"]
5
6     """YOUR CODE HERE FOR Q2.2"""
7     #raise NotImplementedError()
8     fun_obj = LogisticRegressionLoss()
9     optimizer = GradientDescentLineSearchProxL1(10)
10
11    model = linear_models.LogRegClassifier(fun_obj, optimizer)
12    model.fit(X, y)
13

```

¹Harder to insert in `LATEX` than you'd like; turns out there are some drawbacks to using software written in 1978.

```

14     train_err = utils.classification_error(model.predict(X), y)
15     print(f"LogReg Training error: {train_err:.3f}")
16     val_err = utils.classification_error(model.predict(X_valid), y_valid)
17     print(f"LogReg Validation error: {val_err:.3f}")
18     print(f"# nonZeros: {np.sum(model.w != 0)}")
19     print(f"# function evals: {optimizer.num_evals}")

```

Answer:

$\lambda = 0.01$: (1)0(2)0.072(3)89(4)158
 $\lambda = 0.1$: (1)0(2)0.060(3)81(4)236
 $\lambda = 1$: (1)0(2)0.052(3)71(4)107
 $\lambda = 10$: (1)0.050(2)0.090(3)29(4)14

2.3 L0-Regularization [8 points]

The class `LogisticRegressionLossL0` in `fun_obj.py` contains part of the code needed to implement the *forward selection* algorithm, which approximates the solution with L0-regularization,

$$f(w) = \sum_{i=1}^n [\log(1 + \exp(-y_i w^T x_i))] + \lambda \|w\|_0.$$

The class `LogRegClassifierForwardSel` in `linear_models.py` will use a loss function object and an optimizer to perform a forward selection to approximate the best feature set. The `for` loop in its `fit()` method is missing the part where we fit the model using the subset `selected_new`, then compute the score and updates the `min_loss` and `best_feature`. Modify the `for` loop in this code so that it fits the model using only the features `selected_new`, computes the score above using these features, and updates the variables `min_loss` and `best_feature`, as well as `self.total_evals`. Hand in your updated code. Using this new code with $\lambda = 1$, report the training error, validation error, number of features selected, and total optimization steps.

Note that the code differs slightly from what we discussed in class, since we're hard-coding that we include the first (bias) variable. Also, note that for this particular case using the L0-norm with $\lambda = 1$ is using the Akaike Information Criterion (AIC) for variable selection.

Also note that, for numerical reasons, your answers may vary depending on exactly what system and package versions you are using. That is fine.

Answer:

```

1  """YOUR CODE HERE FOR Q2.3"""
2  w_init = np.zeros(selected_with_j.sum())
3  w, fs, gs, ws = self.optimize(w_init, X[:, selected_with_j], y)
4  self.total_evals += self.optimizer.num_evals
5
6  f_global, g_global = self.global_loss_fn.evaluate(w, X[:, selected_with_j], y)
7  if f_global < min_loss:
8      min_loss = f_global
9      best_feature = j

```

Training error: 0.000

Validation error: 0.018

Number of features selected: 24

Number of optimization steps: 138 350

2.4 Discussion [4 points]

In a short paragraph, briefly discuss your results from the above. How do the different forms of regularization compare with each other? Can you provide some intuition for your results? No need to write a long essay, please!

Answer: L2 uses all the features, but the number of features selected decreased in L1 and decreased even more in L0. L0 regularization seems to be slower than the other two as it requires many optimization steps. In addition, looking at the validation errors for each regularization, having more features seems to help avoiding over-fitting.

2.5 $L^{\frac{1}{2}}$ regularization [8 points]

Previously we've considered L2- and L1- regularization which use the L2 and L1 norms respectively. Now consider least squares linear regression with “ $L^{\frac{1}{2}}$ regularization” (in quotation marks because the “ $L^{\frac{1}{2}}$ norm” is not a true norm):

$$f(w) = \frac{1}{2} \sum_{i=1}^n (w^T x_i - y_i)^2 + \lambda \sum_{j=1}^d |w_j|^{1/2}.$$

Let's consider the case of $d = 1$ and assume there is no intercept term being used, so the loss simplifies to

$$f(w) = \frac{1}{2} \sum_{i=1}^n (wx_i - y_i)^2 + \lambda \sqrt{|w|}.$$

Finally, let's assume the very special case of $n = 2$, where our 2 data points are $(x_1, y_1) = (1, 2)$ and $(x_2, y_2) = (0, 1)$.

1. Plug in the dataset values and write the loss in a simplified form, without a \sum .

Answer: $f(w) = \frac{1}{2}(w-2)^2 + \lambda\sqrt{|w|}$

2. If $\lambda = 0$, what is the solution, i.e. $\arg \min_w f(w)$?

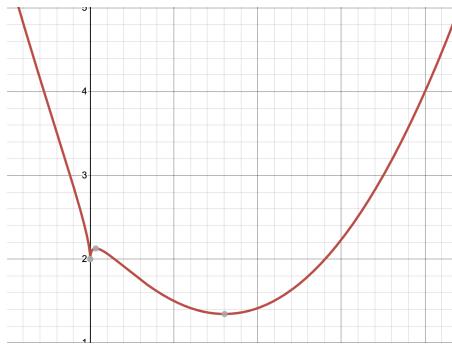
Answer: $w = 2$

3. If $\lambda \rightarrow \infty$, what is the solution, i.e., $\arg \min_w f(w)$?

Answer: $w = 0$

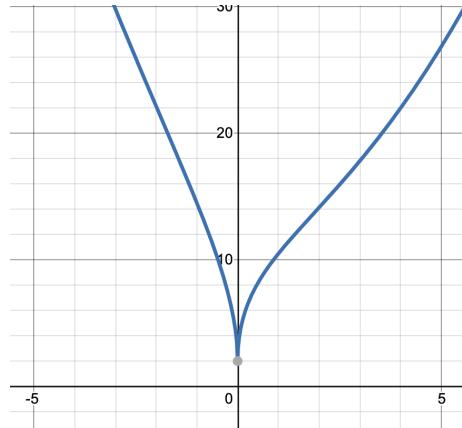
4. Plot $f(w)$ when $\lambda = 1$. What is $\arg \min_w f(w)$ when $\lambda = 1$? Answer to one decimal place if appropriate. (For the plotting questions, you can use `matplotlib` or any graphing software, such as <https://www.desmos.com>.)

Answer: $w = 1.605$



5. Plot $f(w)$ when $\lambda = 10$. What is $\arg \min_w f(w)$ when $\lambda = 10$? Answer to one decimal place if appropriate.

Answer: $w = 0$



6. Does $L^{\frac{1}{2}}$ regularization behave more like L1 regularization or L2 regularization when it comes to performing feature selection? Briefly justify your answer.

Answer: It behaves more like L1 since the coefficient is 0 when λ is extremely big.

7. Is least squares with $L^{\frac{1}{2}}$ regularization a convex optimization problem? Briefly justify your answer.

Answer: From the plot when $\lambda = 1$, it is already showing that it is not convex. Therefore, $L^{\frac{1}{2}}$ regularization is not a convex optimization problem.

3 Multi-Class Logistic Regression [32 points]

If you run `python main.py -q 3` the code loads a multi-class classification dataset with $y_i \in \{0, 1, 2, 3, 4\}$ and fits a “one-vs-all” classification model using least squares, then reports the validation error and shows a plot of the data/classifier. The performance on the validation set is ok, but could be much better. For example, this classifier never even predicts that examples will be in classes 0 or 4.

3.1 Softmax Classification, toy example [4 points]

Linear classifiers make their decisions by finding the class label c maximizing the quantity $w_c^T x_i$, so we want to train the model to make $w_{y_i}^T x_i$ larger than $w_{c'}^T x_i$ for all the classes c' that are not y_i . Here c' is a possible label and $w_{c'}$ is row c' of W . Similarly, y_i is the training label, w_{y_i} is row y_i of W , and in this setting we are assuming a discrete label $y_i \in \{1, 2, \dots, k\}$. Before we move on to implementing the softmax classifier to fix the issues raised in the introduction, let’s work through a toy example:

Consider the dataset below, which has $n = 10$ training examples, $d = 2$ features, and $k = 3$ classes:

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{bmatrix}.$$

Suppose that you want to classify the following test example:

$$\tilde{x} = [1 \ 1].$$

Suppose we fit a multi-class linear classifier using the softmax loss, and we obtain the following weight matrix:

$$W = \begin{bmatrix} +2 & -1 \\ +2 & -2 \\ +3 & -1 \end{bmatrix}$$

Under this model, what class label would we assign to the test example? (Show your work.)

Answer:

Take $c = 1$:

$$[+2 \ -1] \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 2 - 1 = 1$$

Take $c = 2$:

$$[+2 \ -2] \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 2 - 2 = 0$$

Take $c = 3$:

$$[+3 \ -1] \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 3 - 1 = 2$$

$c = 3$ gives us the largest value, so we would assign $c = 3$ to the new test example.

3.2 One-vs-all Logistic Regression [7 points]

Using the squared error on this problem hurts performance because it has “bad errors” (the model gets penalized if it classifies examples “too correctly”). In `linear_models.py`, complete the class named `LogRegClassifierOneVsAll`

that replaces the squared loss in the one-vs-all model with the logistic loss. [Hand in the code and report the validation error.](#)

Answer:

```

1 # NOTE: make sure that you use {-1, 1} labels y for logistic regression,
2 #       not {0, 1} or anything else.
3 for i in range(k):
4     ytmp = y.copy().astype(float)
5     ytmp[y == i] = 1
6     ytmp[y != i] = -1
7     w, fs, gs, ws = LogRegClassifier.optimize(self, W[i], X, ytmp)
8     W[i] = w
9
10 self.W = W

```

Validation error is 0.070.

3.3 Softmax Classifier Gradient [7 points]

Using a one-vs-all classifier can hurt performance because the classifiers are fit independently, so there is no attempt to calibrate the columns of the matrix W . As we discussed in lecture, an alternative to this independent model is to use the softmax loss, which is given by

$$f(W) = \sum_{i=1}^n \left[-w_{y_i}^T x_i + \log \left(\sum_{c'=1}^k \exp(w_c^T x_i) \right) \right],$$

Show that the partial derivatives of this function, which make up its gradient, are given by the following expression:

$$\frac{\partial f}{\partial W_{cj}} = \sum_{i=1}^n x_{ij} [p(y_i = c | W, x_i) - \mathbb{1}(y_i = c)],$$

where...

- $\mathbb{1}(y_i = c)$ is the indicator function (it is 1 when $y_i = c$ and 0 otherwise)
- $p(y_i = c | W, x_i)$ is the predicted probability of example i being class c , defined as

$$p(y_i = c | W, x_i) = \frac{\exp(w_c^T x_i)}{\sum_{c'=1}^k \exp(w_{c'}^T x_i)}$$

$$\begin{aligned}
\text{Answer: } \frac{\partial f}{\partial W_{cj}} &= \sum_{i=1}^n \left[-x_{ij} * \mathbb{1}(y_i = c) + \frac{1}{\sum_{c'=1}^k \exp(w_{c'}^T x_i)} * \exp(w_c^T x_i) * x_{ij} \right] \\
&= \sum_{i=1}^n \left[-x_{ij} * \mathbb{1}(y_i = c) + \frac{1}{\sum_{c'=1}^k \exp(w_{c'}^T x_i)} * \exp(w_c^T x_i) * x_{ij} \right] \\
&= \sum_{i=1}^n \left[x_{ij} \left(\frac{\exp(w_c^T x_i)}{\sum_{c'=1}^k \exp(w_{c'}^T x_i)} - \mathbb{1}(y_i = c) \right) \right] \\
&= \sum_{i=1}^n x_{ij} [p(y_i = c | W, x_i) - \mathbb{1}(y_i = c)]
\end{aligned}$$

3.4 Softmax Classifier Implementation [8 points]

Inside `linear_models.py`, you will find the class `MulticlassLogRegClassifier`, which fits W using the softmax loss from the previous section instead of fitting k independent classifiers. As with other linear models, you must implement a function object class in `fun_obj.py`. Find the class named `SoftmaxLoss`. Complete these classes and their methods. [Submit your code and report the validation error.](#)

Hint: You may want to use `check_correctness()` to check that your implementation of the gradient is correct.

Hint: With softmax classification, our parameters live in a matrix W instead of a vector w . However, most optimization routines (like `scipy.optimize.minimize` or our `optimizers.py`) are set up to optimize with respect to a vector of parameters. The standard approach is to “flatten” the matrix W into a vector (of length kd , in this case) before passing it into the optimizer. On the other hand, it’s inconvenient to work with the flattened form everywhere in the code; intuitively, we think of it as a matrix W and our code will be more readable if the data structure reflects our thinking. Thus, the approach we recommend is to reshape the parameters back and forth as needed. The skeleton code of `SoftmaxLoss` already has lines reshaping the input vector w into a $k \times d$ matrix using `np.reshape`. You can then compute the gradient using sane, readable code with the W matrix inside `evaluate()`. You’ll end up with a gradient that’s also a matrix: one partial derivative per element of W . Right at the end of `evaluate()`, you can flatten this gradient matrix into a vector using `g.reshape(-1)`. If you do this, the optimizer will be sending in a vector of parameters to `SoftmaxLoss`, and receiving a gradient vector back out, which is the interface it wants – and your `SoftmaxLoss` code will be much more readable, too. You may need to do a bit more reshaping elsewhere, but this is the key piece.

Hint: A naïve implementation of `SoftmaxLoss.evaluate()` might involve many for-loops, which is fine as long as the function and gradient calculations are correct. However, this method might take a very long time! This speed bottleneck is one of Python’s shortcomings, which can be addressed by employing pre-computing and lots of vectorized operations. However, it can be difficult to convert your written solutions of f and g into vectorized forms, so you should prioritize getting the implementation to work correctly first. One reasonable path is to first make a correct function and gradient implementation with lots of loops, then (if you want) pulling bits out of the loops into meaningful variables, and then thinking about how you can compute each of the variables in a vectorized way. Our solution code doesn’t contain any loops, but the solution code for previous instances of the course actually did; it’s totally okay for this course to not be allergic to Python `for` loops the way Danica is.²

Answer:

```
1 # softmax loss evaluate function
2 def evaluate(self, w, X, y):
3     w = ensure_1d(w)
4     y = ensure_1d(y)
5
6     n, d = X.shape
7     k = len(np.unique(y))
8
9     """YOUR CODE HERE FOR Q3.4"""
10    g = np.zeros((k,d))
11    W = w.reshape((k, d))
12
13    #LOSS
14    f = 0
15    for i in range(n):
```

²Reading the old solution with loops *probably* isn’t why I was sick the last week....

```

16     sum = 0
17     for c in range(k):
18         sum += math.exp(W[c].T @ X[i])
19     f += -(W[y[i]].T @ X[i]) + math.log(sum)
20
21 #GRADIENT
22 for c in range(k):
23     for j in range(d):
24         sum = 0
25         for i in range(n):
26             x_ij = X[i][j]
27             denominator = 0
28             for l in range(k):
29                 denominator += math.exp(W[l].T @ X[i])
30             numerator = math.exp(W[c].T @ X[i])
31             probability = numerator / denominator
32             indicator_function = 1 if y[i] == c else 0
33             sum += x_ij * (probability - indicator_function)
34         g[c][j] = sum
35 g = g.reshape(-1)
36
37 return f, g
38
39 #fit and predict from MulticlassLogRegClassifier
40 def fit(self, X, y):
41     """YOUR CODE HERE FOR Q3.4"""
42
43     n, d = X.shape
44     y_classes = np.unique(y)
45     k = len(y_classes)
46
47     assert set(y_classes) == set(range(k)) # check labels are {0, 1, ..., k-1}
48
49     # quick check that loss_fn is implemented correctly
50     self.loss_fn.check_correctness(np.zeros(k*d), X, y)
51
52     # Initial guesses for weights
53     W = np.zeros([k, d])
54     W = W.reshape(-1)
55     w, fs, gs, ws = LogRegClassifier.optimize(self, W, X, y)
56
57     W = w.reshape(k, d)
58     self.W = W
59
60 def predict(self, X_hat):
61     """YOUR CODE HERE FOR Q3.4"""
62     return np.argmax(X_hat @ self.W.T, axis=1)

```

Validation error is 0.008.

3.5 Comparison with scikit-learn [2 points]

Compare your results (training error and validation error for both one-vs-all and softmax) with scikit-learn's `LogisticRegression`, which can also handle multi-class problems. For one-vs-all, set `multi_class='ovr'`; for softmax, set `multi_class='multinomial'`. Since your comparison code above isn't using regularization, set `C` very large to effectively disable regularization. Again, set `fit_intercept` to `False` for the same reason as above (there is already a column of 1's added to the data set).

Answer:

For OneVsAll:

Our model reported a training error of 0.084 and a validation error of 0.070.

SciKit Learn's model reported a training error of 0.084 and a validation error of 0.070.

Our model performed exactly as well as SciKit's model.

For SoftMax:

Our model reported a training error of 0.000 and a validation error of 0.008.

SciKit's model reported a training error of 0.010 and a validation error of 0.000.

In this case, I would say that SciKit's model performed a bit better than ours, because there was a higher training error. A training error of 0 could indicate overfitting.

3.6 Cost of Multi-Class Logistic Regression [4 points]

Assume that we have

- n training examples.
- d features.
- k classes.
- t testing examples.
- T iterations of gradient descent for training.

Also assume that we take X and form new features Z using Gaussian RBFs as a non-linear feature transformation.

1. In $O()$ notation, what is the cost of training the softmax classifier with gradient descent?

Answer: $O(T(nk + n^2k^2) + n^2d)$

2. What is the cost of classifying the t test examples?

Answer: $O(ntd + ntk)$

Hint: you'll need to take into account the cost of forming the basis at training (Z) and test (\tilde{Z}) time. It will be helpful to think of the dimensions of all the various matrices.

4 Very-Short Answer Questions [18 points]

Answer each of the following questions in a sentence or two.

1. Suppose that a client wants you to identify the set of “relevant” factors that help prediction. Should you promise them that you can do this?

Answer: We don't know how "relevant" those factors should be, so we might not be able to find the factors that client wants.

2. What is a setting where you would use the L1-loss, and what is a setting where you would use L1-regularization?

Answer:

L1-loss when there are outlier data points

L1-regularization when there are irrelevant features

3. Among L0-regularization, L1-regularization, and L2-regularization: which yield convex objectives? Which yield unique solutions? Which yield sparse solutions?

Answer:

Convex:L1, L2

Unique:L2

Sparse:L0, L1

4. What is the effect of λ in L1-regularization on the sparsity level of the solution? What is the effect of λ on the two parts of the fundamental trade-off?

Answer: As λ increases, the sparsity level of the solution and the training increase whereas the approximation error decreases.

5. Suppose you have a feature selection method that tends not to generate false positives, but has many false negatives (it misses relevant variables). Describe an ensemble method for feature selection that could improve the performance of this method.

Answer: Apply the method to bootstrap samples, and only take features selected in all bootstrap samples

6. Suppose a binary classification dataset has 3 features. If this dataset is “linearly separable”, what does this precisely mean in three-dimensional space?

Answer: There is a hyperplane which is a flat two-dimensional subspace, and the points in two different sides form two different class.

7. When searching for a good w for a linear classifier, why do we use the logistic loss instead of just minimizing the number of classification errors?

Answer: Minimizing classification errors is the 0-1 loss function, so it is non-convex and hard to minimize in terms of w . Logistic loss is convex, so it's easier to minimize.

8. What is a disadvantage of using the perceptron algorithm to fit a linear classifier?

Answer: The perceptron algorithm can only classify the sets of vectors which are linearly separable.

9. How does the hyper-parameter σ affect the shape of the Gaussian RBFs bumps? How does it affect the fundamental tradeoff?

Answer: The hyper-parameter σ affects the width of the bump. As the σ decreases, the training error decreases and the approximation error increases.