Homework #1 A

Spring 2020, CSE 446/546: Machine Learning
Dino Bektesevic

Collaborated: Conor Sayers, Joachim Moeyenes, Jessica Birky, Leah Fulmer

Conceptual Questions *[10 points]*

The answers to these questions should be answerable without referring to external materials. Briefly justify your answers with a few words.

a. *[2 points]* Suppose that your estimated model for predicting house prices has a large positive weight on 'number of bathrooms'. Does it implies that if we remove the feature "number of bathrooms" and refit the model, the new predictions will be strictly worse than before? Why?
If number of bathrooms is independent feature our predictions would be worse. But as we discussed in class, often it is hte case that number of bathrooms is really not an independent feature in the feature set, but instead the weights are split between several different factors that directly or indirectly measure the number of bathrooms.

b. *[2 points]* Compared to L2 norm penalty, explain why a L1 norm penalty is more likely to result in a larger number of 0s in the weight vector or not?
In class we said L1 penalty leads to sparse outputs.

c. *[2 points]* In at most one sentence each, state one possible upside and one possible downside of using the following regularizer: $\sum_i |w_i|^{0.5}$)

d. *[1 points]* True or False: If the step-size for gradient descent is too large, it may not converge.
True. When jumps occur with a step-size much greater than the characteristic step length of a function we can end up constantly jumping to higher values and further away from the minimum.

e. *[2 points]* In your own words, describe why SGD works.
Gradient descent effectively works by starting from a random point and then calculating gradient at that point with respect to all features in the dataset in order to pick the direction of the next point it will step to. Computing true gradients of the sum of squared residuals with respect to features is very costly with datasets containing a lot of points for even moderate number of features. Stochastic gradient descent computes the gradient using a randomly selected point, or a set of points, instead of computing the true gradient using all points. This obviously works for relatively monotonic functions since what we have effectively done was approximating the true function with a line or a hyperplane.

f. *[2 points]* In at most one sentence each, state one possible advantage of SGD (stochastic gradient descent)over GD (gradient descent) and one possible disadvantage of SGD relative to GD.
SGD is much less computationally demanding, but might take a longer time to converge to minimum value than GD might. GD on the other hand virtually guarantees that fewer steps will need to be taken to converge to minimum compared to SGD.

Convexity and Norms *[10 points]*

A.1. A norm $||\cdot||$ over R^n is defined by the properties:

(a) non-negative: $||x|| \geq 0 \forall x \in \mathbb{R}^n \iff x = 0$,

(b) absolute scalability: $||ax|| = |a|||x|| \forall a \in \mathbb{R}$ and $x \in \mathbb{R}^n$,

(c) triangle inequality: $||x + y|| \leq ||x|| + ||y|| \forall x, y \in \mathbb{R}^n$

a. Show that $f(x) = (\sum_{i=1}^{n} |x_i|)$ is a norm. (Hint: begin by showing that $|a + b| \leq |a| + |b| \forall a, b \in \mathbb{R}$.

b. *[2 points]* Show that $g(x) = (\sum_{i}^{n} = \mathbf{1}|x_i|^{1/2})_2$ is not a norm. (Hint: it suffices to find two points in $n = 2$ dimensions such that the triangle inequality does not hold.)

Context: norms are often used in regularization to encourage specific behaviors of solutions. If we define $||x||_p := (\sum_{i}^{n} = \mathbf{1}|x_i|^p)^{1/p}$ then one can show that $||x||^p$ is a norm for all $p \geq 1$. The important cases of $p = 2$ and $p = 1$ correspond to the penalty for ridge regression and the lasso, respectively.

A 2.*[3 points]* A set $A \subseteq \mathbb{R}^n$ is convex if $\lambda x + (1-\lambda)y \in A \forall x, y \in A$ and $\lambda \in [0,1]$. For each of the grey-shaded sets (I-III), state whether each one is convex, or state why it is not convex using any of the points a,b,c,d in your answer.

A 3.*[4 points]* We say a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex on a set $A \iff (\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \forall x, y \in A$ and $\lambda \in [0, 1]$. For each of the grey-colored functions (I-III), state whether each one is convex on the given interval or state why not with a counter example using any of the points a,b,c,d in your answer

a. Function in panel I on [a,c]:

b. Function in panel II on [a,c]:

c. Function in panel III on [a,d]:

d. Function in panel III on [c,d]:

Lasso *[45 points]*

Given $\lambda > 0$ and data $(x_1, y_1), ..., (x_n, y_n)$, the Lasso is the problem of solving

$$\arg \min_{w \in \mathbb{R}^d, b \in \mathbb{R}^n} \sum_{i=1}^{n} (x_i^T w + b - y_i)^2 + \lambda \sum_{j=1}^{d} |w_j|$$

λ is a regularization tuning parameter. For the programming part of this homework, you are required to implement the coordinate descent method of Algorithm 1 that can solve the Lasso problem. You may use common computing packages (such as NumPy or SciPy), but do not use an existing Lasso solver (e.g., of scikit-learn).

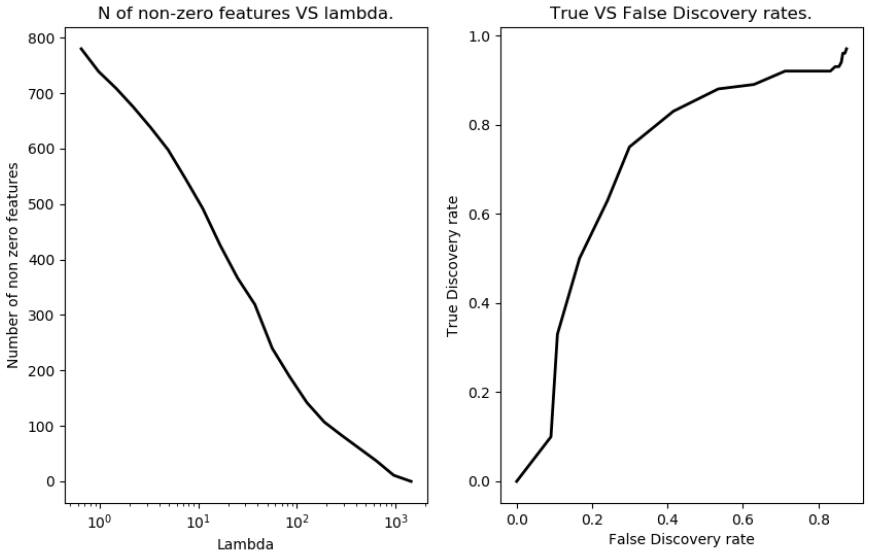
A4. We will first try out your solver with some synthetic data. A benefit of the Lasso is that if we believe many features are irrelevant for predicting y, the Lasso can be used to enforce a sparse solution, effectively differentiating between the relevant and irrelevant features. Suppose that $x \in \mathbb{R}^d, y \in \mathbb{R}, k < d$, and pairs of data (x_i, y_i) for $i = 1, \dots n$ are generated independently according to the model $y_i = w^T x_i + \epsilon_i$ where

$$h(z) = \begin{cases} j/k & \text{if } j \in \{1, \dots, k\} \\ 0 & \text{otherwise} \end{cases}$$

where $\epsilon_i \approx \mathcal{N}(0, \sigma^2)$ is some Gaussian noise (in the model above $b = 0$). Note that since $k < d$, the features $k + 1$ through d are unnecessary (and potentially even harmful) for predicting y. With this model in mind, let $n = 500, d = 1000, k = 100$, and $\sigma = 1$. Generate some data by choosing $x_i \in \mathbb{R}^d$, where each component is drawn from a $\mathcal{N}(0, 1)$ distribution and y_i generated as specified above.

a. *[10 points]* With your synthetic data, solve multiple Lasso problems on a regularization path, starting at $\lambda_{\max}$ where 0 features are selected and decreasing λ by a constant ratio (e.g., 1.5) until nearly all the features are chosen. In plot 1, plot the number of non-zeros as a function of λ on the x-axis (Tip: use log scale).

b. *[10 points]* For each value of λ tried, record values for false discovery rate (FDR) (number of incorrect non zeros in $\widehat{w}$/total number of non zeros in $\widehat{w}$) and true positive rate (TPR) (number of correct non zeros in $\widehat{w}/k$). In plot 2, plot these values with the x-axis as FDR, and the y-axis as TPR and note that in an ideal situation we would have an (FDR,TPR) pair in the upper left corner, but that can always trivially achieve $(0, 0)$ and $(d - kd, 1)$.
For both a and b parts of the problem we have the following graph:



c. *[5 points]* Comment on the effect of λ in these two plots.
No features are selected in the first step, i.e. close to $\lambda_{\max}$. Successive iterations relax regularization penalties so more and more features are selected. Simultaneously TPR increases very rapidly initially and then tapers off the rapid growth. This occurs because we quickly learn the most important features after which adding more features just doesn't add that much more valuable information to the model. Simultaneously our FDR skyrockets because the additional features over-specify our model. Eventually TPR jumps to nearly one as, I assume, we just fit all the points.

Conclusions provided courtesy of my brain, figures provided courtesy of the following code:

```
import matplotlib.pyplot as plt
import numpy as np


def descent(x, y, lambd, tolerance, initW=None):
    """Preforms coordinate descent Lasso algorithm on the given data.

    Parameters
    ----------
    x : `np.array`
        The feature space. A matrix with n rows of feature vectors, each with d
        features.
    y : `np.array`
        A column vector of n model values.
    lambd: `float`
        Regularization parameter.
    tolerance: `float`
        Convergence is achieved when the absolute value of minimal difference
        of old weights and the new weights estimates is less than tolerance.
    initW: `np.array`
        Initial weights, a vector of d feature weights.

    Returns
    -------
    w: `np.array`
        New feature weights estimates.
    """
    n, d = x.shape

    if initW is None:
        w = np.zeros(d)
    else:
        w = initW

    # precalculate values used in the loop in advance
    squaredX =  2.0 * x**2

    # ensure convergence is not met on first loop
    convergeCriterion = tolerance + 1
    while convergeCriterion > tolerance:
        # not optimal test, but fast
        oldMax = w.max()
        deltas = []

        # Algorithm 1 implementation
        b = np.mean(y - np.dot(x, w))
        for k in range(d):
            xk = x[:, k]
            ak = squaredX[:,k].sum()

            # ck sum must ignore k-th dimension so we set it to zero and use
            # dot product. This matches the definition of w too, so we can
            # leave it wth zero value unless smaller than -lambda or bigger
            # than lambda anyhow.
            w[k] = 0
            delta = 0
            ck = 2.0 * np.dot(xk, y - (b + np.dot(x, w)))

            if ck < -lambd:
                delta = (ck + lambd) / ak
                w[k] = delta
            elif ck > lambd:
                delta = (ck - lambd) / ak
                w[k] = delta

            deltas.append(delta)
        # Find maximum difference between iterations
        convergeCriterion = abs(oldMax-max(deltas))
    return w


def generate_data(n, d, k, sigma):
    """Generates i.i.d. samples of the model:
        y_i = w^T x_i + eps
    where
        w_j = j/k if j in {1,...,k}
        w_j = 0 otherwise
    and epsilon is random Gussian noise with the given sigma and X are also
    drawn from a Normal distribution with sigma 1.

    Parameters
    ----------
    n : `int`
        Number of samples drawn at random from the model.
    d : `int`
        Dimensionality of the feature space.
    k : `int`
        Cutoff point after which elements of w are zero.

    Returns
    -------
```

```
    x : `np.array`
        n-by-d sized array of data.
    y : `np.array`
        Vector of n model values.
    """
    # gaussian noise and data
    eps = np.random.normal(0, sigma**2, size=n)
    x = np.random.normal(size=(n, d))

    # weights
    w = np.arange(1, d + 1) / k
    w[k:] = 0

    # labels
    y = np.dot(x, w) + eps
    return x, y


def lambda_max(x, y):
    """The smallest value of regularization parameter lambda for which the
    w is entirely zero.

    Parameters
    ----------
    x : `np.array`
        The feature space. A matrix with n rows of feature vectors, each with d
        features.
    y : `np.array`
        A column vector of n model values.

    Returns
    -------
    lambda: `float`
        Smallest lambda for which w is entirely zero.
    """
    return np.max(2 * np.abs(np.dot((y - np.mean(y)), x)))


def A4setup(n=500, d=1000, k=100, sigma=1):
    """Creates data as instructed by A4 problem, calculates the smallest value
    of regularization parameter for which w is zero and returns data parameters,
    data and calculated lambda.

    Parameters
    ----------
    n : `int`, optional
        Number of samples drawn at random from the model. Default: 500
    d : `int`, optional
        Dimensionality of the feature space. Default: 1000
    k : `int`, optional
        Cutoff point after which elements of w are zero. Default: 100
    sigma: `float`, optional
        STD of the noise Gaussian distribution that is added to the data (see
        generate_data). Default: 1.

    Returns
    -------
    x : `np.array`
        The feature space. A matrix with n rows of feature vectors, each with d
        features.
    y : `np.array`
        A column vector of n model values.
    maxLambda: `float`
        Lambda for which w is zero everywhere.
    params: `dict`
        Dictionary of parameters used to create the data (n, d, k and sigma).
    """
    params = {'n': n, 'd': d, "k": k, "sigma": sigma}
    x, y = generate_data(n, d, k, sigma)
    maxLambda = lambda_max(x, y)
    return x, y, maxLambda, params


def plot(ax, x, y, label="", xlabel="", ylabel="", title="", xlog=True, lc='black', lw=5):
    """Plots a line on given axis.

    Parameters
    ----------
    ax: `matplotlib.pyplot.Axes`
       Axis to plot on.
    x: `np.array`
       X axis values
    y: `np.array`
       Y axis values
    label: `str`, optional
       Line label
    xlabel: `str`, optional
       X axis label
    ylabel: `str`, optional
       Y axis label
    title: `str`, optional
       Axis title
    xlog : `bool`, optional
       X axis scaling will be logarithmic
    lc : `str`, optional
       Line color
    lw: `int` or `float`, optional
       Line width

    Returns
    ------
    ax: `matplotlib.pyplot.Axes`
        Modified axis.
    """
    ax.set_title(title)
    ax.plot(x, y, label=label, color=lc, linewidth=lw)
    ax.set_xlabel(xlabel)
    ax.set_ylabel(ylabel)
    if xlog:
        ax.set_xscale('log')
```

```
    return ax


def A4(nIter=20, tolerance=0.001):
    """Sets the data up as instructed by problem A4 and runs coordinate
    descent Lasso algorithm nIter times, each time decreasing regularization
    parameter lambda by a factor of 1.5.
    Plots the number of non-zero-features against used lambda and false vs
    true discovery rates.

    Displays plots.

    Parameters
    ----------
    nIter: `int`, optional
        Number of different regularization parameter iterations to run. Default
        is 20.
    tolerance: `float`, optional
        Coordinate descent tolerance, sets convergence criteria (see descent).
        Default: 0.001.
    """
    x, y, lambd, params = A4setup()
    w = np.zeros(params['d'])
    k = params['k']

    lambdas, numNonZeros, fdrs, tprs = [], [], [], []
    for i in range(nIter):
        # Find w_hat
        w = descent(x, y, lambd, tolerance, initW=w)

        nonZeros = np.count_nonzero(w)
        correctNonZeros = np.count_nonzero(w[:k])
        incorrectNonZeros = np.count_nonzero(w[k+1:])

        try:
            fdrs.append(incorrectNonZeros/nonZeros)
        except ZeroDivisionError:
            fdrs.append(0)
        tprs.append(correctNonZeros/k)

        lambdas.append(lambd)
        numNonZeros.append(nonZeros)

        lambd /= 1.5

    fig, axes = plt.subplots(1, 2, figsize=(10, 6))
    plot(axes[0], lambdas, numNonZeros, xlabel="Lambda",
         ylabel="Number of non zero features",
         title="N of non-zero features VS lambda.")
    plot(axes[1], fdrs, tprs, xlabel="False Discovery rate",
         ylabel="True Discovery rate", title="True VS False Discovery rates.",
         xlog=False)
    plt.show()


if __name__ == "__main__":
    A4()
```

A5. Now we put the Lasso to work on some real data. Download the training data set "crime-train.txt" and the test data set "crime-test.txt" from the website under Homework 2. Store your data in your working directory and read in the files with:

```
import pandas as pd
df_train = pd.read_table("crime-train.txt")
df_test = pd.read_table("crime-test.txt")
```

The data consist of local crime statistics for 1,994 US communities. The response y is the crime rate. The name of the response variable isViolentCrimesPerPop, and it is held in the first column of df_train and df_test. There are 95 features. These features include possibly relevant variables such as the size of the police force or the percentage of children that graduate high school. The data have been split for you into a training and test set with 1,595 and 399 entries, respectively.
We'd like to use this training set to fit a model which can predict the crime rate in new communities and evaluate model performance on the test set. As there are a considerable number of input variables, over fitting is a serious issue. In order to avoid this, use the coordinate descent LASSO algorithm you just implemented in the previous problem. Begin by running the LASSO solver with $\lambda = \lambda_{\max}$ defined above. For the initial weights, just use 0. Then, cut λ down by a factor of 2 and run again, but this time pass in the values of $\widehat{w}$ from your $\lambda = \lambda_{\max}$ solution as your initial weights. This is faster than initializing with 0 weights each time. Continue the process of cutting λ by a factor of 2 until the smallest value of λ is less than 0.01. For all plots use a log-scale for the λd dimension.

a. *[4 points]* Plot the number of nonzeros of each solution versus λ

b. *[4 points]* Plot the regularization paths (in one plot) for the coefficients for input variables agePct12t29, pctWSocSec, pctUrban, agePct65up, and householdsize.

c. *[4 points]* Plot the squared error on the training and test data versus λ

d. *[4 points]* Sometimes a larger value of λ performs nearly as well as a smaller value, but a larger value will select fewer variables and perhaps be more interpretable. Inspect the weights (on features) for $\lambda = 30$. Which feature variable had the largest (most positive) Lasso coefficient? What about the most negative? Discuss briefly. A description of the variables in the data set can be found here: `http://archive.ics.uci.edu/ml/machine-learning-databases/communities/communities.names`.

e. *[4 points]* Suppose there was a large negative weight on agePct65up and upon seeing this result, a politician suggests policies that encourage people over the age of 65 to move to high crime areas in an effort to reduce crime. What is the (statistical) flaw in this line of reasoning? (Hint: fire trucks are often seen around burning buildings, do fire trucks cause fire?)

Binary Logistic Regression *[30 points]*

A6. Let us again consider the MNIST dataset, but now just binary classification, specifically, recognizing if a digit is a 2 or 7. Here, let Y= 1 for all the 7's digits in the dataset, and use Y=-1 for 2. We will use regularized logistic regression. Given a binary classification dataset $(x_i, y_i)_{i=1}^n$ for $x_i \in \mathbb{R}$ dand $y_i \in -1, 1$ we showed in class that the regularized negative log likelihood objective function can be written a

$$J(w, b) = \frac{1}{n} \sum_{i=1}^{n} \log 1 + e^{-y_i(b+x_i^T w)} + \lambda ||w||_2^2$$

Note that the offset term b is not regularized. For all experiments, use $\lambda = 10^{-1}$. Let $\mu_i(w, b) = \frac{1}{1+exp(-y_i(b+x_i^T w))}$

a. *[8 points]* Derive the gradients $\nabla w J(w, b)$, $\nabla b J(w, b)$ and give your answers in terms of $\mu_i(w, b)$ (your answers should not contain exponentials)

b. *[8 points]* Implement gradient descent with an initial iterate of all zeros. Try several values of step sizes to find one that appears to make convergence on the training set as fast as possible. Run until you feel you are near to convergence.

   (a) For both the training set and the test, plot $J(w, b)$ as a function of the iteration number (and show both curves on the same plot)

   (b) For both the training set and the test, classify the points according to the rule $sign(b + x_i^T w)$ and plot the misclassification error as a function of the iteration number (and show both curves on the same plot). Note that you are only optimizing on the training set. The $J(w, b)$ and misclassification error plots should be on separate plots.

c. *[7 points]* Repeat (b) using stochastic gradient descent with a batch size of 1. Note, the expected gradient with respect to the random selection should be equal to the gradient found in part (a). Take careful note of how to scale the regularizer.

d. *[7 points]* Repeat (b) using stochastic gradient descent with batch size of 100. That is, instead of approximating the gradient with a single example, use 100. Note, the expected gradient with respect to the random selection should be equal to the gradient found in part (a).