

CM146, Winter 2021
Problem Set 2: Perceptron and regression
Due February 12, 2021 at 11:59 pm

Submission

- Submit your solutions electronically on the course Gradescope site as PDF files.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.

Parts of this assignment are adapted from course material by Andrew Ng (Stanford), Jenna Wiens (UMich) and Jessica Wu (Harvey Mudd).

1 Perceptron [2 pts]

Design (specify θ for) a two-input perceptron (with an additional bias or offset term) that computes the following boolean functions. Assume $T = 1$ and $F = -1$. If a valid perceptron exists, show that it is not unique by designing another valid perceptron (with a different hyperplane, not simply through normalization). If no perceptron exists, state why.

(a) OR

(b) XOR

2 Logistic Regression [10 pts]

Consider the objective function that we minimize in logistic regression:

$$J(\theta) = - \sum_{n=1}^N [y_n \log h_{\theta}(\mathbf{x}_n) + (1 - y_n) \log (1 - h_{\theta}(\mathbf{x}_n))]$$

(a) Find the partial derivatives $\frac{\partial J}{\partial \theta_j}$.

(b) Find the partial second derivatives $\frac{\partial^2 J}{\partial \theta_j \partial \theta_k}$ and show that the Hessian (the matrix \mathbf{H} of second derivatives with elements $H_{jk} = \frac{\partial^2 J}{\partial \theta_j \partial \theta_k}$) can be written as $\mathbf{H} = \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) \mathbf{x}_n \mathbf{x}_n^T$.

(c) Show that J is a convex function and therefore has no local minima other than the global one.

Hint: A function J is convex if its Hessian is positive semi-definite (PSD), written $\mathbf{H} \succeq 0$. A matrix is PSD if and only if

$$\mathbf{z}^T \mathbf{H} \mathbf{z} \equiv \sum_{j,k} z_j z_k H_{jk} \geq 0.$$

for all real vectors \mathbf{z} .

3 Maximum Likelihood Estimation [15 pts]

Suppose we observe the values of n independent random variables X_1, \dots, X_n drawn from the same Bernoulli distribution with parameter θ ¹. In other words, for each X_i , we know that

$$P(X_i = 1) = \theta \quad \text{and} \quad P(X_i = 0) = 1 - \theta.$$

Our goal is to estimate the value of θ from these observed values of X_1 through X_n .

For any hypothetical value $\hat{\theta}$, we can compute the probability of observing the outcome X_1, \dots, X_n if the true parameter value θ were equal to $\hat{\theta}$. This probability of the observed data is often called the *likelihood*, and the function $L(\theta)$ that maps each θ to the corresponding likelihood is called the *likelihood function*. A natural way to estimate the unknown parameter θ is to choose the θ that maximizes the likelihood function. Formally,

$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(\theta).$$

- (a) Write a formula for the likelihood function, $L(\theta) = P(X_1, \dots, X_n; \theta)$. Your function should depend on the random variables X_1, \dots, X_n and the hypothetical parameter θ . Does the likelihood function depend on the order in which the random variables are observed ?
- (b) Since the log function is increasing, the θ that maximizes the *log likelihood* $\ell(\theta) = \log(L(\theta))$ is the same as the θ that maximizes the likelihood. Find $\ell(\theta)$ and its first and second derivatives, and use these to find a closed-form formula for the MLE.
- (c) Suppose that $n = 10$ and the data set contains six 1s and four 0s. Write a short program `likelihood.py` that plots the likelihood function of this data for each value of $\hat{\theta}$ in $\{0, 0.01, 0.02, \dots, 1.0\}$ (use `np.linspace(...)` to generate this spacing). For the plot, the x-axis should be θ and the y-axis $L(\theta)$. Scale your y-axis so that you can see some variation in its value. Include the plot in your writeup (there is no need to submit your code). Estimate $\hat{\theta}_{MLE}$ by marking on the x-axis the value of $\hat{\theta}$ that maximizes the likelihood. Does the answer agree with the closed form answer ?
- (d) Create three more likelihood plots: one where $n = 5$ and the data set contains three 1s and two 0s; one where $n = 100$ and the data set contains sixty 1s and forty 0s; and one where $n = 10$ and there are five 1s and five 0s. Include these plots in your writeup, and describe how the likelihood functions and maximum likelihood estimates compare for the different data sets.

¹This is a common assumption for sampling data. So we will denote this assumption as iid, short for Independent and Identically Distributed, meaning that each random variable has the same distribution and is drawn independent of all the other random variables

4 Implementation: Polynomial Regression [40 pts]

In this exercise, you will work through linear and polynomial regression. Our data consists of inputs $x_n \in \mathbb{R}$ and outputs $y_n \in \mathbb{R}, n \in \{1, \dots, N\}$, which are related through a target function $y = f(x)$. Your goal is to learn a linear predictor $h_\theta(x)$ that best approximates $f(x)$. But this time, rather than using `scikit-learn`, we will further open the “black-box”, and you will implement the regression model!

code and data

- `code`: CS146_Winter2021_PS2.ipynb
 - `data`: regression_train.csv, regression_test.csv
-

Please use your `@g.ucla.edu` email id to access the code and data. Similar to *HW-1*, copy the colab notebook to your drive and make the changes. Mount the drive appropriately and copy the shared data folder to your drive to access via colab. For colab usage demo, check out the Discussion recordings for Week 2 in CCLE. The notebook has marked blocks where you need to code.

===== *TODO : START* =====

===== *TODO : END* =====

Note: For the questions requiring you to complete a piece of code, you are expected to copy-paste your code as a part of the solution in the submission pdf. Tip: If you are using L^AT_EX, check out the Minted package ([example](#)) for code highlighting.

This is likely the first time that many of you are working with `numpy` and matrix operations within a programming environment. For the uninitiated, you may find it useful to work through a `numpy` tutorial first.² Here are some things to keep in mind as you complete this problem:

- If you are seeing many errors at runtime, inspect your matrix operations to make sure that you are adding and multiplying matrices of compatible dimensions. Printing the dimensions of variables with the `X.shape` command will help you debug.
- When working with `numpy` arrays, remember that `numpy` interprets the `*` operator as element-wise multiplication. This is a common source of size incompatibility errors. If you want matrix multiplication, you need to use the `dot` function in Python. For example, `A*B` does element-wise multiplication while `dot(A,B)` does a matrix multiply.
- Be careful when handling `numpy` vectors (rank-1 arrays): the vector shapes $1 \times N$, $N \times 1$, and N are all different things. For these dimensions, we follow the the conventions of `scikit-learn`’s `LinearRegression` class³. Most importantly, unless otherwise indicated (in the code documentation), both column and row vectors are rank-1 arrays of shape N , not rank-2 arrays of shape $N \times 1$ or shape $1 \times N$.

Visualization [2 pts]

²Try out SciPy’s tutorial (http://wiki.scipy.org/Tentative_NumPy_Tutorial), or use your favorite search engine to find an alternative. Those familiar with Matlab may find the “Numpy for Matlab Users” documentation (http://wiki.scipy.org/NumPy_for_Matlab_Users) more helpful.

³http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

It is often useful to understand the data through visualizations. For this data set, you can use a scatter plot to visualize the data since it has only two properties to plot (x and y).

- (a) Visualize the training and test data using the `plot_data(...)` function. What do you observe? For example, can you make an educated guess on the effectiveness of linear regression in predicting the data? [2 pts]

Linear Regression [23 pts]

Recall that linear regression attempts to minimize the objective function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^N (h_{\boldsymbol{\theta}}(\mathbf{x}_n) - y_n)^2.$$

In this problem, we will use the matrix-vector form where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_D \end{pmatrix}$$

and each instance $\mathbf{x}_n = (1, x_{n,1}, \dots, x_{n,D})^T$.

In this instance, the number of input features $D = 1$.

Rather than working with this fully generalized, multivariate case, let us start by considering a simple linear regression model:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} = \theta_0 + \theta_1 x_1$$

The Colab notebook contains the skeleton code for the class `PolynomialRegression`. Objects of this class can be instantiated as `model = PolynomialRegression(m)` where m is the degree of the polynomial feature vector where the feature vector for instance n , $(1, x_{n,1}, x_{n,1}^2, \dots, x_{n,1}^m)^T$. Setting $m = 1$ instantiates an object where the feature vector for instance n , $(1, x_{n,1})^T$.

- (b) Note that to take into account the intercept term (w_0), we can add an additional “feature” to each instance and set it to one, e.g. $x_{i,0} = 1$. This is equivalent to adding an additional first column to \mathbf{X} and setting it to all ones [2 pts].

Modify `PolynomialRegression.generate_polynomial_features(...)` to create the matrix \mathbf{X} for a simple linear model.

- (c) Before tackling the harder problem of training the regression model, complete `PolynomialRegression.predict(...)` to predict \mathbf{y} from \mathbf{X} and $\boldsymbol{\theta}$. [3 pts]
- (d) One way to solve linear regression is through gradient descent (GD).

Recall that the parameters of our model are the θ_j values. These are the values we will adjust to minimize $J(\boldsymbol{\theta})$.

$$J(\boldsymbol{\theta}) = \sum_{n=1}^N (h_{\boldsymbol{\theta}}(\mathbf{x}_n) - y_n)^2$$

In gradient descent, each iteration performs the update

$$\theta_j \leftarrow \theta_j - 2\alpha \sum_{n=1}^N (h_{\boldsymbol{\theta}}(\mathbf{x}_n) - y_n) x_{n,j} \quad (\text{simultaneously update } \theta_j \text{ for all } j).$$

With each step of gradient descent, we expect our updated parameters w_j to come closer to the parameters that will achieve the lowest value of $J(\boldsymbol{\theta})$. [10 pts]

- As we perform gradient descent, it is helpful to monitor the convergence by computing the cost, *i.e.*, the value of the objective function J . Complete `PolynomialRegression.cost(...)` to calculate $J(\boldsymbol{\theta})$.

If you have implemented everything correctly, then the following code snippet should return 40.234.

```
train_data = load_data(train_path)
model = PolynomialRegression()
model.coef_ = np.zeros(2)
model.cost(train_data.X, train_data.y)
```

- Next, implement the gradient descent step in `PolynomialRegression.fit_GD(...)`. The loop structure has been written for you, and you only need to supply the updates to $\boldsymbol{\theta}$ and the new predictions $\hat{y} = h_{\boldsymbol{\theta}}(\mathbf{x})$ within each iteration.

We will use the following specifications for the gradient descent algorithm:

- We run the algorithm for 10,000 iterations.
- We terminate the algorithm earlier if the value of the objective function is unchanged across consecutive iterations.
- We will use a fixed learning rate.
- Experiment with different values of learning rate $\eta = 10^{-6}, 10^{-5}, 10^{-3}, 0.0168$ and make a table of the coefficients, number of iterations until convergence (this number will be 10,000 if the algorithm did not converge in a smaller number of iterations) and the final value of the objective function. How do the coefficients compare? How quickly does each algorithm converge? Do you observe something strange when you run with $\eta = 0.0168$? Explain the observation and causes.

(e) In class, we learned that the closed-form solution to linear regression is

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Using this formula, you will get an exact solution in one calculation: there is no “loop until convergence” like in gradient descent. [4 pts]

- Implement the closed-form solution `PolynomialRegression.fit(...)`.

- What is the closed-form solution coefficients? How do the coefficients and the cost compare to those obtained by GD? How quickly does the algorithm run compared to GD?
- (f) Finally, set a learning rate η for GD that is a function of k (the number of iterations) (use $\eta_k = \frac{1}{1+k}$) and converges to the same solution yielded by the closed-form optimization (minus possible rounding errors). Update `PolynomialRegression.fit_GD(...)` with your proposed learning rate. How many iterations does it take the algorithm to converge with your proposed learning rate? [4 pts]

Polynomial Regression[15 pts]

Now let us consider the more complicated case of polynomial regression, where our hypothesis is

$$h_{\theta}(\mathbf{x}) = \boldsymbol{\theta}^T \phi(\mathbf{x}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_m x^m.$$

- (g) Recall that polynomial regression can be considered as an extension of linear regression in which we replace our input matrix \mathbf{X} with

$$\Phi = \begin{pmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T \end{pmatrix},$$

where $\phi(x)$ is a function such that $\phi_j(x) = x^j$ for $j = 0, \dots, m$.

Update `PolynomialRegression.generate_polynomial_features(...)` to create an $m + 1$ dimensional feature vector for each instance. [4 pts]

- (h) Given N training instances, it is always possible to obtain a “perfect fit” (a fit in which all the data points are exactly predicted) by setting the degree of the regression to $N - 1$. Of course, we would expect such a fit to generalize poorly. In the remainder of this problem, you will investigate the problem of overfitting as a function of the degree of the polynomial, m . To measure overfitting, we will use the Root-Mean-Square (RMS) error, defined as

$$E_{RMS} = \sqrt{J(\boldsymbol{\theta})/N},$$

where N is the number of instances.⁴

Why do you think we might prefer RMSE as a metric over $J(\boldsymbol{\theta})$?

Implement `PolynomialRegression.rms_error(...)`. [4 pts]

- (i) For $m = 0, \dots, 10$ (where m is the order of the polynomial transformation applied to features), use the closed-form solver to determine the best-fit polynomial regression model on the training data, and with this model, calculate the RMSE on both the training data and the test data. Generate a plot depicting how RMSE varies with model complexity (polynomial degree) – you should generate a single plot using `plot_erms(...)` function, and include this plot in your writeup. Which degree polynomial would you say best fits the data? Was there evidence of under/overfitting the data? Use your plot to justify your answer. [7 pts]

⁴Note that the RMSE as defined is a biased estimator. To obtain an unbiased estimator, we would have to divide by $n - k$, where k is the number of parameters fitted (including the constant), so here, $k = m + 1$.