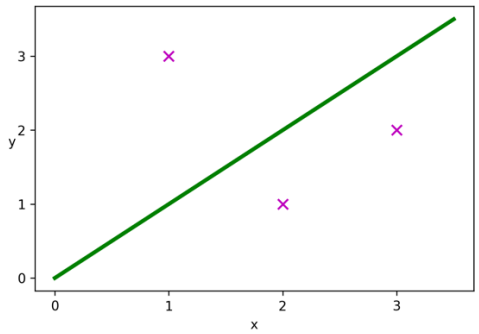
CS305 Exercise 4

Task 1: Linear Regression

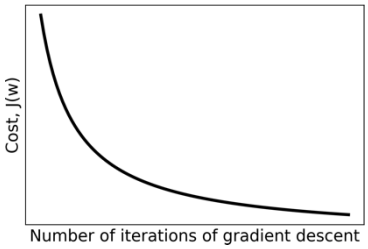
As shown in the image below, suppose we have three training data points (1,3), (2,1), and (3,2) and, using linear regression, we learn a set of parameters $\boldsymbol{w}$ = (0, 1).

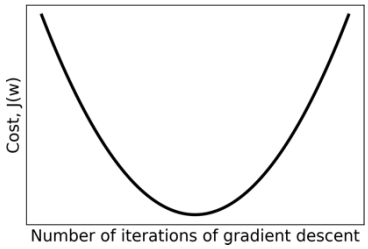


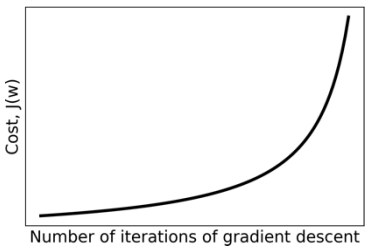
What is the cost $J(\boldsymbol{w})$ associated with this set of parameters $\boldsymbol{w}$ = (0, 1) on the training data?

In linear regression, if the cost function achieves a value of zero on a set of training data points, what property do the training data points have?

When gradient descent executes, it typically performs many iterations, updating parameters with each iteration. Circle the figure below that represents the relationship between the number of iterations of gradient descent and the cost function as gradient descent executes, assuming gradient descent is working as desired.







If you observe the cost function increasing with each iteration of gradient descent, it would be reasonable to assume:

(A) the learning rate α is too small
(B) the learning rate α is about right
(C) the learning rate α is too large

Suppose you have a large dataset on which you are running regularized linear regression with batch gradient descent and you observe that it is taking a long time to complete training of your model. Describe two ways that gradient descent might be modified in order to speed things up.

When using regularized linear regression, using too large of a value for the regularization parameter λ is likely to lead to underfitting or overfitting?

When using regularized linear regression, using too small of a value for the regularization parameter λ is likely to lead to underfitting or overfitting?

When using regularized linear regression, what strategy might you use for identifying a good value for the regularization parameter λ?

Is a machine learning algorithm more likely to perform better on training data or on validation data?

Is a machine learning algorithm more likely to perform better on validation data or on test data?

Previously, when we studied the *k*-Nearest-Neighbors algorithm, we considered it as an approach for solving *classification* problems. How would you adapt the *k*-Nearest-Neighbors algorithm so that it could be used to solve *regression* problems?

Task 2: Cross-Validation

In the validation approaches we have used thus far, we have generally split data into two separate groups, one used for training and one used for validation. However, there are other cross-validation approaches.

In *k*-fold cross-validation, the data are split into *k* equally sized groups. Then, there are *k* rounds of training/validation:

- The first of the *k* groups is designated as validation data and the other *k*-1 groups are used as training data.
- The second of the *k* groups is designated as validation data and the other *k*-1 groups are used as training data.
- The third of the *k* groups is designated as validation data and the other *k*-1 groups are used as training data.
- …
- The last of the *k* groups is designated as validation data and the other *k*-1 groups are used as training data.

Thus, each of the *k* groups (and each data point) is used exactly once for validation purposes. The final reported validation score is the average score over the *k* training/validation rounds.

For example, in 2-fold cross-validation, data are split into two equal sized groups. First, one group is used for training and the other for validation. Then they are switched, so that the training group becomes the validation group and vice versa. The final validation score is the average of the 2 training/validation rounds.

10-fold cross-validation is common, where data are split into 10 equal sized groups and 10 rounds of training/validation occur.

When *k*=*n*, where *n* is the number of data points, there are *n* rounds of training/validation, where each round uses *n*-1 examples for training and a single example for validation. This is known as "leave-one-out cross-validation" (LOOCV).

What is one advantage of using *k*-fold cross-validation as opposed to simply splitting the data into two separate groups, one for training and one for validation?

What is one disadvantage of using *k*-fold cross-validation as opposed to simply splitting the data into two separate groups, one for training and one for validation?

Suppose we apply linear regression to some training data, and when we use the learned model on test data we find that it performs poorly. We wonder if the data is non-linear, so we try adding different new sets of features to our data, including various higher-order polynomial combinations of our original features. For each set of added higher-order features, we train a model using training data and then evaluate the model's performance on test data. Some of the added features result in models with good performance on the test data. Why is this approach problematic?

Download the Jupyter Notebook for Exercise 4 from the course website. Open the Notebook in your web browser and work through it. As you work through the Notebook, answer the following questions.

Task 3: Predicting Movie Revenue

How many movies are there in the dataset, i.e., what is *n*? How many features does each movie have, i.e., what is *d*?

Looking at your plot, about how much revenue is predicted for a movie with a budget of $300 million?

What was the score when using only one feature (movie budget) for linear regression? What was the score when using multiple features for linear regression? Did the score improve when more features were used?

What *three* features have the most significant (lowest) p-values (effectively 0.0), i.e., contribute most to determining a movie's revenue?

Task 4: How Much Does a Diamond Cost?

What is the R^2 coefficient of your linear regression model using the testing data?

What *four* features have the most significant (lowest) p-values (effectively 0.0), i.e., contribute most to determining a diamond's price?

What is the predicted price of the above diamond with clarity I1? What is the predicted price of the above diamond with clarity IF?

CS305 Exercise 4 Final Page

Name(s): _____

In the *TIME* column, please estimate the time you spent on this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

PART	TIME	SCORE
Exercise		