Introduction to Machine Learning
Problem Set: Linear Regression and Gradient Descent

Summer 2021

1. **Residual**

   Is the following statement *True* or *False*? Explain your answer.

   > The residual of a regression model, $y - \hat{y}$, is random error with no pattern or systematic trend. There is no machine learning model that can "explain" any part of the residual.

2. **Regression metrics**. For the following questions, your answers can be in the form of a range, like "$-\infty$ to ∞" or "≤ 0", along with a brief explanation.

   Suppose you have training data and test data sampled from the data-generating process

   $$y = A + Bx + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

   where A and B are real valued scalar constants.

   You fit an ordinary least squares linear regression model to the training data. Finally, you compute R2 for the fitted model.

   (a) What is the range of values you could get for R2 on your training data?
   (b) What is the range of values you could get for R2 on your test data?

   Answer the same questions, but assuming $\sigma = 0$:

   (a) What is the range of values you could get for R2 on your training data?
   (b) What is the range of values you could get for R2 on your test data?

   Answer the same questions again, but if the *test data* was actually sampled from

   $$y = C + Dx + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

   where C and D are real valued scalar constants. Also assume $\sigma = 0$.

   (a) What is the range of values you could get for R2 on your training data?
   (b) What is the range of values you could get for R2 on your test data?

3. **Linear basis function regression**.

You have labeled data $(x_i, y_i), i = 1, \ldots, n$ and you want to fit an exponential model of the form,

$$\hat{y_i} = \sum_{j=0}^{d} w_j e^{-jx_i}$$

where x_i and y_i are scalars. You will use a linear basis function regression.

(a) Given training data $((x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5), (x_6, y_6))$, write out the entries of the design matrix Φ you would use to fit the model above, for $d = 2$.

(b) Suppose you have training data loaded into a `numpy` array `x` in Python, and you also have a variable with the value of `d` (not necessarily 2 for this part). Write Python code to create the design matrix and save it in another variable, `X`.

Try to minimize your use of **`for`** loops - in fact, you can try to write an answer that is a single line of code!

4. **Gradient descent**.

For this question you will change some parameters in the "Gradient descent in depth" notebook, re-run the notebook with the new parameters,and answer questions about the results. You do not have to write any new code, and you should not submit any code.

(Copy the relevant output images to a regular document, answer the questions there, and submit that document - don't submit the Colab notebook with all the gradient descent code.)

(a) Re-run the "Descent path" section with three different learning rates: `lr = 0.0002`, `lr = 0.002`, and `lr = 0.02` (and leave other parameters at their default settings). For each learning rate,

i. Show the plot of coefficient value vs. iteration, and the plot of the descent path on the MSE contour.

ii. What is the estimate of w after 50 iterations?

iii. Describe whether the gradient descent diverges, converges within 50 iterations, or starts to converge but does not get very close to the optimum value within 50 iterations.

(b) Re-run the "Stochastic gradient descent" section with `lr=0.1` and `n=1`, then with `lr=0.01` and `n=10`, and finally with `lr = 0.001` and `n = 100` (and leave the other parameters at their default settings).

(Note: in this question, we are primarily observing the effect of changing the mini-batch size, n. We are varying the learning rate only so that the learning rate per sample is constant as we increase the mini-batch size.)

For each,

i. Show the plot of coefficient value vs. iteration, and the plot of the descent path on the MSE contour.

ii. Comment on the descent path. Does it go smoothly toward the optimal solution?

5. **Linear regression on the Advertising data**.

   Please refer to the homework notebook posted on the class site.