

Exercises

§9.1,9.3,9.4: Preparing Features

Python: There is an optional starter script that supports Python-based problem(s) on this sheet. Use it as scaffolding: read the comments, complete the lines marked **TODO**, and compare the output with the mathematics requested in the problem(s).

Problem 1 A feature has training values 2, 4, 6, 8. Using denominator P : (a) compute the mean, (b) compute the standard deviation $s = \sqrt{\frac{1}{P} \sum_p (x_p - \mu)^2}$, (c) standardize all four values, (d) verify that the standardized values have mean 0 and mean squared value 1.

Problem 2 A regression model is $\hat{y} = 5 + 2x$. Define $z = (x - 10)/4$. Rewrite the same prediction rule in the form $\hat{y} = \tilde{w}_0 + \tilde{w}_1 z$. Verify that the two formulas give the same prediction when $x = 14$.

Problem 3 For $X = \begin{bmatrix} -1 & -100 \\ -1 & 100 \\ 1 & -100 \\ 1 & 100 \end{bmatrix}$, compute $H = (2/4)X^T X$. Find its eigenvalues and condition number. Then

divide the second feature by 100 and repeat. Explain why gradient descent should be easier in the rescaled coordinates.

Problem 4 Suppose three unordered categories are A, B, C. (a) Write their three-column one-hot encoding. (b) Explain why using all three indicator columns together with an intercept creates a linearly dependent design matrix. (c) Using A as the reference category, write the regression prediction $\hat{y} = w_0 + w_B B + w_C C$ for each of the three categories.

Problem 5 A training feature has $\mu_{\text{train}} = 20$ and $s_{\text{train}} = 5$. Two test observations are 15 and 35. Standardize the test observations correctly. Why would recomputing a mean and standard deviation from the two test observations be conceptually wrong?

Problem 6 Show that standardizing features changes Euclidean squared distance from $\sum_j (x_j - a_j)^2$ to $\sum_j \frac{(x_j - a_j)^2}{s_j^2}$. Interpret this as a weighted Euclidean distance. Which features receive greater geometric weight: those with large s_j or those with small s_j ?

Problem 7 Use Python to fit a standardizer on

```
x_train = np.array([[1., 100.],
                    [2., 120.],
                    [3., 140.],
                    [4., 160.]])
```

and transform

```
x_test = np.array([[5., 180.],
                   [6., 200.]])
```

using the training mean and training scale. Report **mu**, **scale**, **Z_train**, and **Z_test**.

Problem 8 Explain separately how feature scaling affects: (a) an unregularized linear regression model, (b) gradient descent used to fit that model, (c) k -means clustering, (d) PCA, (e) a linear model with a coefficient penalty. The answer should distinguish parameterization, optimization geometry, and model geometry.