Section 04: Data Normalization, Common ML Errors, Vector Calculus

1. Data Normalization/Standardization

Sometimes, our features have very different ranges of values. This is not ideal and can lead to numerical issues (e.g., overflow) and optimization difficulties.

There are two ways to take care of this issue. One is called data normalization, and the other is called data standardization. Sometimes these terms are used interchangeably, but it is important to understand the difference.

Below, $x_i^{(j)}$ represents the value of the i^{th} feature of the j^{th} data point.

1.1. Data Standardization

Data standardization is the task of transforming each feature in our dataset to have mean 0 and variance 1. The typical way to do this is using the Z-Score, which is defined as below:

$$\tilde{x_i}^{(j)} = \frac{x_i^{(j)} - \mu_i}{\sigma_i}$$

Where μ_i is the mean of each feature and σ_i is the standard deviation of each feature, which are empirically calculated from the data.

Question: what should you do when $\sigma_i = 0$ for some i?

1.2. Data Normalization

Data normalization refers to the task of rescaling each feature in our dataset to have range [0, 1].

One such method to achieve this is min-max scaling:

$$\tilde{x_i}^{(j)} = \frac{x_i^{(j)} - x_i^{min}}{x_i^{max} - x_i^{min}}$$

Where x_i^{min}, x_i^{max} are the minimum and maximum values of feature i in our dataset, respectively.

When training and evaluating your model, you should calculate the parameters for your normalization or standardization function on the training set **ONLY**!

In other words, if we were using Z-Score, we'd calculate our μ_i, σ_i on the training set, and use those same values when standardizing our validation/test data. The same applies to normalization methods, such as min-max scaling, where we want to determine x_i^{min}, x_i^{max} from training data. This will likely mean that we may get some values outside [0, 1] after rescaling new data, but, given that our training dataset is large enough, this shouldn't be much of an issue (so unseen data likely won't be wildly outside of [0, 1]).

Question 1: Should we always choose x_i^{min} and x_i^{max} based on train data? Can we sometimes do better? Think about cases when we have some underlying information about data.

Question 2: When can values outside of [0, 1] range in test set cause issues?

2. Regularization

Just a reminder, don't ever regularize your bias term. This term doesn't add any complexity to the model (since it just shifts), so we'd like it to take on any value that best fits our training data.

3. Vector Calculus

3.1. Definitions

Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $g : \mathbb{R}^n \to \mathbb{R}^m$. The **gradient** of f(with respect to x) evaluated at x is the vector of partial derivatives:

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

The **Jacobian** of g(with respect to x) evaluated at x is the matrix of partial derivatives:

$$\nabla_x g(x) = \begin{bmatrix} \frac{\partial g_1(x)}{\partial x_1} & \cdots & \frac{\partial g_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m(x)}{\partial x_1} & \cdots & \frac{\partial g_m(x)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla_x^T g_1(x) \\ \vdots \\ \nabla_x^T g_m(x) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Sometimes the Jacobian is denoted by $J_g(x)$, but we use $\nabla_x g(x)$ to highlight that the Jacobian is nothing more than the generalization of the gradient to functions which have a vector output.

The **Hessian** of f(with respect to x) evaluated at x is the matrix of partial derivatives:

$$\nabla_x^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

Sometimes the Hessian is denoted by $H_f(x)$, but we use $\nabla_x^2 f(x)$ to highlight that the Hessian is the Jacobian of the gradient of f.

1 Let $f(x_1, x_2) = x_1^2 + e^{x_1 x_2} + 2\log(x_2)$. What are the gradient and the Hessian of f?

2 Note that $\nabla_x f : \mathbb{R}^n \to \mathbb{R}^n$. What is the Jacobian of $\nabla_x f$?

3.2. Estimation

What the gradient and Jacobian at a point do is to express how the output of a function changes when the input is changed by a small amount. Thus, they can be used to approximate the values of a function close to the point at which they are evaluated. Let's see how we can do this for one variable. Let $f : \mathbb{R} \to \mathbb{R}$:

$$\frac{df}{dx}(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon} \Leftrightarrow \frac{df}{dx}(x) \approx \frac{f(x+\epsilon) - f(x)}{\epsilon} \Leftrightarrow f(x+\epsilon) \approx f(x) + \epsilon \frac{df}{dx}(x)$$

Let us now extend this to multiple dimensions and derive the definition of the gradient starting from this approximation view point. Suppose we have a function $f : \mathbb{R}^n \to \mathbb{R}$ and we want to determine how the function changes around a point $x \in \mathbb{R}^n$. First we will determine how the function changes when we slightly vary its first coordinate:

$$f(x_1 + \epsilon_1, \ldots, x_n) \approx f(x_1, \ldots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, \ldots, x_n)$$

Now, let us slightly vary the first two coordinates:

$$\begin{aligned} f(x_1+\epsilon_1, x_2+\epsilon_2, \dots, x_n) &\approx f(x_1, x_2+\epsilon_2, \dots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2+\epsilon_2, \dots, x_n) \\ &\approx f(x_1, x_2, \dots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \dots, x_n) + \\ &+ \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \dots, x_n) + \epsilon_1\epsilon_2 \frac{\partial f}{\partial x_2}\frac{\partial f}{\partial x_1}(x_1, x_2, \dots, x_n) \\ &\approx f(x_1, x_2, \dots, x_n) + \epsilon_1 \frac{\partial f}{\partial x_1}(x_1, x_2, \dots, x_n) + \epsilon_2 \frac{\partial f}{\partial x_2}(x_1, x_2, \dots, x_n) \end{aligned}$$

where we eliminate the term where $\epsilon_1\epsilon_2$ because it would be very small compared to the others. Repeating the process for all n dimensions we obtain the approximation:

$$f(x_1+\epsilon_1, \dots, x_n+\epsilon_n) \approx f(x_1, \dots, x_n) + \sum_{i=1}^{n} \epsilon_i \frac{\partial f}{\partial x_i}(x_1, x_2, \dots, x_n)$$

Let $\epsilon = [\epsilon_1, \dots, \epsilon_n]^T$ and $x = [x_1, \dots, x_n]^T$, then we can rewrite the above as:

$$f(x+\epsilon) \approx f(x) + \nabla_x f(x)^T \epsilon$$

1 The gradient $\nabla_x f(x)$ offers the best linear approximation of f around the point x. What does the Jacobian of a function $g : \mathbb{R}^n \to \mathbb{R}^m$ offer?

2 If we use the gradient and the Hessian of $f : \mathbb{R}^n \to \mathbb{R}$, what type of an approximation for the function f around a point x can we create.

3 Consider the function $f(x_1, x_2) = 2 + 0.2(x_1 - 3)^2 + 0.2(x_2 - 3)^2$ which is graphed below. The pink plane is the tangent plane for the point $x = (4, 4)$ and it represents the graph of the best linear approximation of f around the point x. What is the function describing the tangent plane:

4 One thing to note is that the linear approximation becomes very poor once we move away from x. Suppose we want a better approximation. For this purpose, we can use the Hessian as explained in part 2. Write down this approximation for an arbitrary x. How good would this approximation be?

5 Draw the gradient on the picture. Describe what happens to the values of the approximation of f if we move from x in directions d_1, d_2, d_3 for which $\nabla_x f(x)^T d_1 > 0, \nabla_x f(x)^T d_2 < 0, \nabla_x f(x)^T d_3 = 0$? Can the same conclusions be drawn about the function of f?

3.3. Algebra

Let $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$, . Below is a list of important gradient properties:

- **Gradient of constant:** $\nabla_x c = 0 \in \mathbb{R}^n$ for a constant $c \in \mathbb{R}^n$.
- **Linearity:** $\nabla_x(\alpha f + \beta g)(x) = \alpha \nabla_x f(x) + \beta \nabla_x g(x)$ for a scalars $\alpha, \beta \in \mathbb{R}$.
- **Product rule**: $\nabla_x(fg)(x) = \nabla_x f(x) \cdot g(x) + \nabla_x g(x) \cdot f(x)$.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^m \to \mathbb{R}^k$, $l : \mathbb{R}^m \to \mathbb{R}$. Below is a list of important Jacobian properties:

- **Jacobian of constant:** $\nabla_x c = 0 \in \mathbb{R}^{n \times m}$ for a constant $c \in \mathbb{R}^n$.
- **Linearity:** $\nabla_x(\alpha f + \beta g)(x) = \alpha \nabla_x f(x) + \beta \nabla_x g(x)$ for a scalars $\alpha, \beta \in \mathbb{R}$.
- **Product rule**: $\nabla_x(f^T g)(x) = [\nabla_x f(x)]^T g(x) + [\nabla_x g(x)]^T f(x)$.
- **Chain rule**: $\nabla_x(h \circ g)(x) = \nabla_{g(x)} h(g(x)) \nabla_x g(x)$ and $\nabla_x(l \circ g)(x) = \left[[\nabla_{g(x)} l(g(x))]^T \nabla_x g(x)\right]^T$.

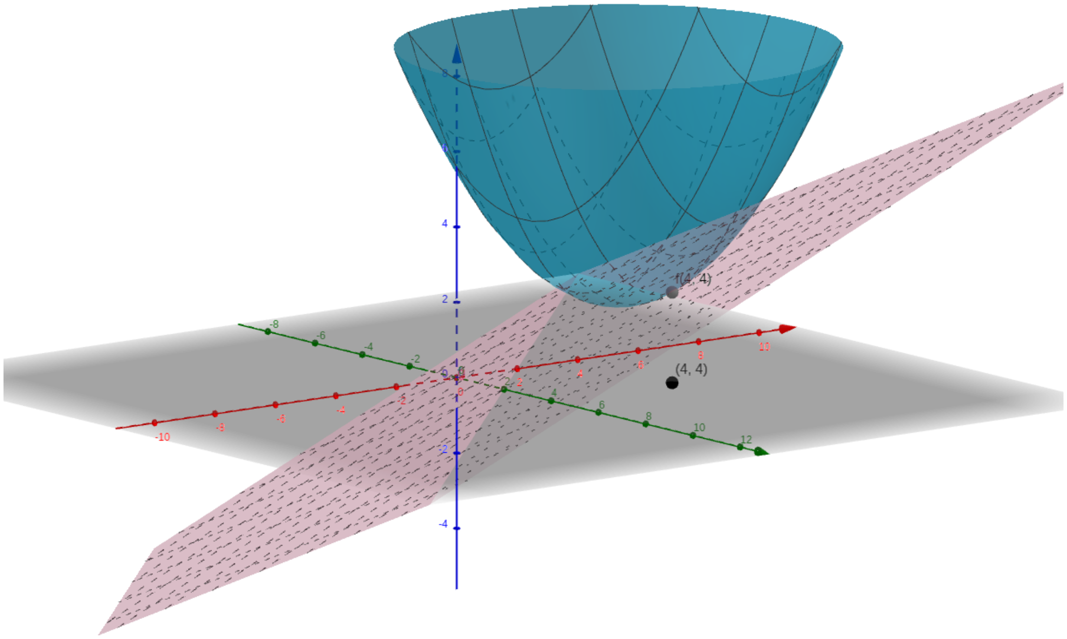

Figure 1: Graph of the function f and the tangent plane.

Questions:

1 Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = v^T x$ for $v \in \mathbb{R}^n$. Using the definition of the gradient, write out $\nabla_x f(x)$ and specify its dimensions.

2 Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be $f(x) = x$. Using the definition of the Jacobian, write out $\nabla_x f(x)$ and specify its dimensions.

3 Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be $f(x) = Ax$ for $A \in \mathbb{R}^{m \times n}$. Using the definition of the Jacobian, write out $\nabla_x f(x)$ and specify its dimensions.

4 Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = \alpha v^T x + \beta w^T x$ where $\alpha, \beta \in \mathbb{R}$ and $v, w \in \mathbb{R}^n$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

5 Let $f : \mathbb{R}^n \to \mathbb{R}$ be $f(x) = x^T Ax$ and $A \in \mathbb{R}^{n \times n}$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.

6 With f defined as in the previous part, what is the Hessian of f. Only use previously proven facts and recall that the Hessian is the Jacobian of the gradient.

7 Let $f : \mathbb{R}^m \to \mathbb{R}$ be $f(x) = (Ax - y)^T W(Ax - y)$ and $A \in \mathbb{R}^{m \times n}, W \in \mathbb{R}^{n \times n}, y \in \mathbb{R}^n$. Using the properties at the beginning of the section and previous results, write out $\nabla_x f(x)$.