3. Linear Neural Networks for Regression - Conceptual Exercises

Gabriel Romão

January 20, 2025

3.1. Linear Regression

1. Assume that we have some data $x_1, \ldots, x_n \in \mathbb{R}$. Our goals is to find a constant b such that $\sum_i (x_i - b)^2$.

a. Find an analytic solution for the optimal value of b.

$$
\begin{aligned}
f(b) &= \sum (x_i - b)^2 \\
f(b) &= \sum (x_i^2 - 2x_i b + b^2) \\
f(b) &= \sum x_i^2 - 2b \sum x_i + nb^2 \\
f'(b) &= -2 \sum x_i + 2nb \\
-2 \sum x_i + 2nb &= 0 \\
\cancel{2}nb &= \cancel{2} \sum x_i \\
b &= \frac{\sum x_i}{n}
\end{aligned}
$$

b. How does this problem and its solution relate to the normal distribution?

b is the mean of the data.

2. Prove that the affine functions that can be expressed by $\mathbf{x}^\top \mathbf{w} + b$ are equivalent to linear functions on $(\mathbf{x}, 1)$

The affine function:

$$f(x) = \mathbf{x}^\top \mathbf{w} + b$$

The linear funcion:

$$g(z) = \mathbf{z}^\top \mathbf{v}$$

where:

- $\mathbf{z} = (\mathbf{x}, 1) \in \mathbb{R}$
- $\mathbf{v} = (\mathbf{w}, b) \in \mathbb{R}$

Solving replacing z and v for their contents:

$$g(z) = (\mathbf{x}, 1)^\top (\mathbf{w}, b)$$
$$g(z) = \mathbf{x}^\top \mathbf{w} + b$$

4. Recall that one of the conditions for the linear regression problem to be solvable was that the design matrix $\mathbf{X}^\top \mathbf{X}$ has full rank.

a. What happens if this is not the case?

$\mathbf{X}^\top \mathbf{X}$ will not be invertible, thus the solution will not be unique.

b. How could you fix it? What happens if you add a small amount of coordinate-wise independent Gaussian noise to all entries of $\mathbf{X}$?

This could work if, in the end, the $\mathbf{X}$ is full rank.

$$\mathbf{X} = \mathbf{X} + \mathbf{Z}$$

where:

- $z_{ij} \sim \mathcal{N}(0, 1)$

c. What is the expected value of the design matrix $\mathbf{X}^\top \mathbf{X}$ in this case?

$$\begin{aligned}
\mathbf{X}^\top \mathbf{X} &= (\mathbf{X} + \mathbf{Z})^\top (\mathbf{X} + \mathbf{Z}) \\
&= (\mathbf{X}^\top + \mathbf{Z}^\top)(\mathbf{X} + \mathbf{Z}) \\
&= \mathbf{X}^\top \mathbf{X} + \mathbf{X}^\top \mathbf{Z} + \mathbf{Z}^\top \mathbf{X} + \mathbf{Z}^\top \mathbf{Z}
\end{aligned}$$

d. What happens with stochastic gradient descent when $\mathbf{X}^\top \mathbf{X}$ does not have full rank?

- **Multiple Solutions**: The solution is not unique and may lie in the null space of $\mathbf{X}$.
- **Slow Convergence**: Redundancy among features creates flat regions in the loss function.
- **Unstable Updates**: Updates in poorly constrained directions can be unstable.

Regularization (e.g., Ridge regression), feature selection, or adding noise can help mitigate these problems.

5. Assume that the noise model governing the additive noise ϵ is the exponential distribution. That is, $p(\epsilon) = \frac{1}{2} \exp(-|\epsilon|)$.

a. Write out the negative log-likelihood of the data under the model $-\log P(\boldsymbol{y}|\boldsymbol{X})$.

$$y = \boldsymbol{w}^\top \boldsymbol{x} + b + \epsilon \text{ where } \epsilon \sim Exp(\lambda)$$

The likelihood:

$$
\begin{aligned}
P(\boldsymbol{y}|\boldsymbol{X}) &= \frac{1}{2}\exp(-|y - \boldsymbol{w}^\top \boldsymbol{x} - b|) \\
-\log P(\boldsymbol{y}|\boldsymbol{X}) &= \sum_{i=1}^{n} \log(2) + |y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b|
\end{aligned}
$$