Solutions Week-04: Logistic and Softmax Regression

1. **Sigmoid Calculation:**

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

   - $z = -2$: $\sigma(-2) = \frac{1}{1+e^{2}} \approx \frac{1}{1+7.389} \approx \frac{1}{8.389} \approx 0.119$
   - $z = 0$: $\sigma(0) = \frac{1}{1+e^{0}} = \frac{1}{1+1} = 0.5$
   - $z = 1.5$: $\sigma(1.5) = \frac{1}{1+e^{-1.5}} \approx \frac{1}{1+0.223} \approx \frac{1}{1.223} \approx 0.818$

2. **Simple Logistic Prediction:** $P = \sigma(-2 + 0.5 \times 6) = \sigma(-2+3) = \sigma(1) = \frac{1}{1+e^{-1}} \approx \frac{1}{1+0.3679} \approx 0.731$

3. **Decision Boundary:** Set $P(Y=1|x) = 0.5$. This happens when the input to σ is 0. $3 - 2x = 0$ $2x = 3$ $x = 1.5$

4. **Log Loss (Single Point):** Log Loss $= -[y\log(p) + (1-y)\log(1-p)] = -[1 \times \log(0.85) + 0] = -\log(0.85) \approx -(-0.1625) \approx 0.1625$

5. **Log Loss (Multiple Points):**

   - Point 1: $-[1 \times \log(0.9)] = -\log(0.9) \approx 0.1054$
   - Point 2: $-[0 \times \log(0.2) + 1 \times \log(0.8)] = -\log(0.8) \approx 0.2231$
   - Point 3: $-[1 \times \log(0.6)] = -\log(0.6) \approx 0.5108$

   Total Loss $= 0.1054 + 0.2231 + 0.5108 = 0.8393$ Average Log Loss $= 0.8393/3 \approx 0.2798$

6. **Two-Feature Prediction:** $z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = -1 + (0.4)(2) + (-0.8)(3) = -1 + 0.8 - 2.4 = -2.6$ $P = \sigma(-2.6) = \frac{1}{1+e^{2.6}} \approx \frac{1}{1+13.4637} \approx 0.069$

7. **Cost Function Derivative:** Term $= (y - \hat{y})x_j = (1 - 0.7) \times 2 = 0.3 \times 2 = 0.6$ The positive sign indicates that increasing θ_j will increase the prediction $\hat{y}$ for this point, moving it closer to the true label $y = 1$. The weight θ_j will be increased during the gradient update.

8. **Softmax Logits:** $\mathbf{z} = [1.2, -0.5, 0.8]$ $e^{z} \approx [e^{1.2}, e^{-0.5}, e^{0.8}] \approx [3.320, 0.6065, 2.226]$ Sum $= 3.320 + 0.6065 + 2.226 = 6.1525$

   - $P(\text{Class 1}) = 3.320/6.1525 \approx 0.540$
   - $P(\text{Class 2}) = 0.6065/6.1525 \approx 0.099$
   - $P(\text{Class 3}) = 2.226/6.1525 \approx 0.362$

9. **Softmax Prediction:** The highest probability is 0.540 for Class 1. Therefore, Class 1 is predicted.

10. **Softmax Constant Shift:** Original logits: $[1.2, -0.5, 0.8]$ Subtract max (1.2): $[0, -1.7, -0.4]$ $e^{z} \approx [1, 0.1827, 0.6703]$ Sum $= 1 + 0.1827 + 0.6703 = 1.853$

- $P(\text{Class 1}) = 1/1.853 \approx 0.540$
- $P(\text{Class 2}) = 0.1827/1.853 \approx 0.099$
- $P(\text{Class 3}) = 0.6703/1.853 \approx 0.362$

The probabilities are identical, proving the invariance.

11. **Multi-class Log Loss:** True class is 2, so the true probability vector is $[0, 1, 0]$. Predicted vector is $[0.1, 0.7, 0.2]$. Log Loss $= -\sum_{c=1}^{3} y_c \log(p_c) = -[0 \times \log(0.1) + 1 \times \log(0.7) + 0 \times \log(0.2)] = -\log(0.7) \approx 0.3567$

12. **Parameter Update Interpretation:** Gradient $= -2.5$. This negative value indicates that increasing θ_1 increases the error for this data point. Update: $\Delta\theta_1 = -\eta \times \text{gradient} = -0.1 \times (-2.5) = +0.25$ So, θ_1 will be increased by 0.25.

13. **Effect of Feature Scaling:** Gradient descent converges faster when features are on similar scales. If features have different scales, the parameter updates will be uneven, leading to an oscillating path to the minimum. Scaling ensures a smoother, faster convergence.

14. **Overfitting Prevention:**

    (a) **L1 (Lasso) or L2 (Ridge) Regularization:** Adds a penalty to the cost function for large parameter values.

    (b) **Early Stopping:** Stopping the training process when performance on a validation set starts to degrade.

15. **Softmax with High Confidence:** True label is first class: $[1, 0, 0, 0]$ Predicted: $[0.94, 0.02, 0.03, 0.01]$ Log Loss $= -[1 \times \log(0.94) + 0 + 0 + 0] = -\log(0.94) \approx 0.0619$

16. **Softmax Symmetry:** For binary classification, let logits be z_1 and z_2. $P(Y = 1) = \frac{e^{z_1}}{e^{z_1}+e^{z_2}} = \frac{1}{1+e^{z_2-z_1}}$ Let $z = z_1 - z_2$. Then $P(Y = 1) = \frac{1}{1+e^{-z}} = \sigma(z)$. QED.

17. **Complex Decision Boundary:** $z = 1+2(1)-3(1)+0.5(1\times 1) = 1+2-3+0.5 = 0.5$ $P = \sigma(0.5) = \frac{1}{1+e^{-0.5}} \approx \frac{1}{1+0.6065} \approx 0.622$

18. **Gradient Calculation:** Gradient with respect to θ_0: $(y - \hat{y})x_0 = (1 - 0.3) \times 1 = 0.7$

19. **Regularization Effect:** The L2 regularization term in the gradient is $\alpha\theta_j = 0.1 \times 2.5 = 0.25$ This term would be added to the gradient, encouraging smaller parameter values.

20. **Multi-class Decision:** Predicted class is class 2 (highest probability 0.6). If true class is class 3, the probability of the true class is 0.3.