CSCI 145 Problem 6: Choosing a Loss Is Choosing a Noise Model

July 25, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 6: Choosing a Loss Is Choosing a Noise Model

In lecture, every data point carried the same noise level σ, and Gaussian noise handed us mean squared error. Real measurements are rarely so uniform: a lab scale reads a package to the gram while a shipping estimate is off by kilos, and a payroll record is far tighter than self-reported income. Squared error is easy to mistake for a convention. This problem changes the noise assumption twice, first letting the noise level vary from point to point and then dropping the Gaussian altogether, and watches the loss change each time.

Throughout, f is a hypothesis, the n data points $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$ are independent, and

$$r^{(i)} = y^{(i)} - f(\mathbf{x}^{(i)})$$

is the residual of f on point i, exactly as in lecture.

Part A: Two Sensors, Two Noise Levels

Suppose point 1 comes from a precise sensor, $y^{(1)} = f(\mathbf{x}^{(1)}) + \eta_1$ with $\eta_1 \sim \mathcal{N}(0, \sigma_1^2)$, and point 2 comes from a sloppy one, $y^{(2)} = f(\mathbf{x}^{(2)}) + \eta_2$ with $\eta_2 \sim \mathcal{N}(0, \sigma_2^2)$ and $\sigma_2 \gg \sigma_1$ (picture $\sigma_1 = 1$ and $\sigma_2 = 10$). Following lecture's two-point derivation, write the joint likelihood $L(f)$ as a product of two Gaussian densities with *different* variances, take its negative log, and simplify to

$$-\log L(f) = \log(2\pi\sigma_1\sigma_2) + \frac{(r^{(1)})^2}{2\sigma_1^2} + \frac{(r^{(2)})^2}{2\sigma_2^2}.$$

Now compare the two residual terms. If a small change in f moves both residuals by the same amount, which term changes more, and does that match your intuition about which sensor the fit should listen to?

Part B: Weighted Least Squares

Now give every point its own noise level: $\eta_i \sim \mathcal{N}(0, \sigma_i^2)$ for $i = 1, \ldots, n$, independently. Repeat the negative-log-likelihood computation of Part A for the full dataset and show that maximizing the likelihood over f is equivalent to

$$\arg\min_f \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \big(y^{(i)} - f(\mathbf{x}^{(i)})\big)^2,$$

i.e., *weighted* least squares, where point i carries weight $1/\sigma_i^2$. Then specialize to constant hypotheses $f(\mathbf{x}) = c$, differentiate with respect to c, and show the optimal constant is the noise-weighted average

$$c^\star = \frac{\sum_{i=1}^n y^{(i)}/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2}.$$

Check both results against lecture by setting every σ_i to the same σ: the objective should collapse to mean squared error, and $c^\star$ to the sample mean $\bar{y}$. Then explain in one sentence why a point with small σ_i pulls $c^\star$ toward itself more than a point with large σ_i does.

Part C: Laplace Noise, Absolute Error

Gaussian noise was a choice, not a law of nature. The **Laplace distribution** has density

$$p(\eta) = \frac{1}{2b} \exp\left(-\frac{|\eta|}{b}\right)$$

for a scale parameter $b > 0$. Suppose instead that $y^{(i)} = f(\mathbf{x}^{(i)}) + \eta_i$ with $\eta_i \sim \text{Laplace}(0, b)$ for every i, independently, with the same b throughout. Write the joint likelihood, take its negative log, and show that maximizing the likelihood over f is equivalent to

$$\arg\min_f \sum_{i=1}^{n} \left|y^{(i)} - f(\mathbf{x}^{(i)})\right|,$$

i.e., mean absolute error. Lecture showed that the constant minimizing this objective is the sample median rather than the sample mean, so the same dataset now gets a different fit. Conclude in one sentence: what does "choose a loss function" actually mean?

Where this goes next. The recipe you used three times here (write the likelihood, take a log, turn the argmax into an argmin) is the single tool this course uses to justify every loss function it introduces. It returns in two weeks when Bayes' rule and a likelihood ratio hand us the sigmoid function for logistic regression, and again near the end of the semester when a Gaussian *prior* on the weights, rather than on the noise, turns maximum likelihood into the ridge penalty.