CS109 Lecture 25 Worksheet: Regression and Reinforcement Learning

Beyond Classification: Linear Regression as MLE, Least Squares, and Learning from Rewards

Summer 2026

Lecture 25 Worksheet: Regression and Reinforcement Learning

Topics: the landscape of machine learning tasks, the linear regression model, Gaussian noise, deriving least squares from maximum likelihood, the gradient and gradient ascent, estimating the noise variance, and reinforcement learning as learning from rewards.

How to use this sheet. Work in small groups. A TA and the instructor will circulate to help. We will go over selected solutions on the board. Today's payoff is that *the most famous formula in statistics — "minimize the sum of squared errors" — is not an arbitrary choice.* It falls out of maximum likelihood estimation the moment you assume Gaussian noise. You have had all the pieces since Lecture 9.

Useful facts. $\sigma(z) = \dfrac{1}{1+e^{-z}}$, $\quad \sigma(0.4) = 0.5987$. The Normal PDF is $f(z) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$. $\quad \Phi(2.24) \approx 0.9873$.

Notation (matching the slides). Each datapoint i has a feature vector $x^{(i)}$ (with $x_0^{(i)} = 1$ prepended as the bias) and a real-valued output $y^{(i)}$. The model's prediction is $\hat{y}^{(i)} = \theta^T x^{(i)}$.

Problem 1 (Review of Lecture 22): One Step of Backpropagation

A neural network is evaluated on a single datapoint with input $x = [1,\ 2,\ 0]$ (bias prepended) and true label $y = 1$. The forward pass produced hidden activations $h = [1,\ 0.6,\ 0.2]$ (bias prepended), and the output weights are $\theta^{(2)} = [0,\ 1,\ -1]$.

(a) Compute the weighted sum $z^{(2)}$ into the output neuron and then $\hat{y}$.

(b) What is the log-likelihood of this datapoint?

(c) Compute all three output-layer gradients $\frac{\partial LL}{\partial \theta_j^{(2)}} = h_j\,[y - \hat{y}]$.

(d) Compute $\frac{\partial LL}{\partial \theta_{1,1}^{(1)}} = [y - \hat{y}]\,\theta_1^{(2)} h_1(1 - h_1) x_1$.

(e) **Bridge to today.** Suppose we deleted the sigmoid on the output neuron, so the network emits the raw number $z^{(2)}$ instead of a probability. The label y is now a real number rather than a 0 or a 1. What probability model would you assume for $Y \mid X = x$ now, and why can we no longer use the Bernoulli?

Problem 2: The Landscape of Machine Learning Tasks

For each task below, name which of the five categories from lecture it belongs to — binary classification, multi-class classification, regression, reinforcement learning, or generation — and state what the output random variable Y is (including its type and support).

(a) Predicting whether a patient has heart disease from 13 measurements.

(b) Predicting the number of points a basketball team will score in tonight's game from the opposing team's ELO, whether the game is at home, and the points scored in the last game.

(c) Predicting which of 1000 object categories appears in a photograph.

(d) Predicting the next token given all previous tokens.

(e) Learning to keep a simulated critter alive by moving toward apples and away from poison, given only a score at the end.

(f) In lecture we saw that the output of an LLM is a Categorical distribution over the vocabulary. Given that, explain in one sentence why "generation" is not really a sixth kind of model, but rather a way of *using* one of the others.

Problem 3: The Linear Regression Model

Logistic regression assumed $Y \mid X = x \sim \text{Bern}(\sigma(\theta^T x))$. For a real-valued output we assume instead that the true relationship is linear but corrupted by noise:

$$y^{(i)} = \theta^T x^{(i)} + Z^{(i)}, \qquad Z^{(i)} \sim N(0, \sigma^2), \quad \text{i.i.d.}$$

(a) Using the properties of the Normal distribution from Lecture 9, show that this is equivalent to

$$Y^{(i)} \mid X = x^{(i)} \;\sim\; N\Big(\theta^T x^{(i)},\ \sigma^2\Big).$$

Which two facts about Normals did you use?

(b) In words: what is being assumed about the *errors* the model makes? State three separate assumptions buried in the line above.

(c) Why is it essential that the noise has mean 0? What would happen to the fitted θ_0 if the noise actually had mean 5?

(d) Notice that σ^2 is the same for every datapoint. Give a real dataset where you would expect that assumption to fail.

(e) Compare the two models side by side. Fill in the table:

	Logistic Regression	Linear Regression
Distribution of $Y \mid X = x$		
Range of the prediction		
Function applied to $\theta^T x$		
Log-likelihood of one datapoint		

Problem 4: Deriving Least Squares from Maximum Likelihood

This is the central derivation of the day. We follow the same three steps as always: make an assumption, write the log-likelihood, take derivatives.

(a) Write the likelihood of all n datapoints, $L(\theta) = \prod_{i=1}^{n} f\left(y^{(i)} \mid x^{(i)}; \theta\right)$, substituting the Normal PDF with mean $\theta^T x^{(i)}$.

(b) Take the log and simplify to show

$$LL(\theta) = \sum_{i=1}^{n} \log\left[\frac{1}{\sigma\sqrt{2\pi}}\right] - \frac{1}{2}\left(\frac{y^{(i)} - \theta^T x^{(i)}}{\sigma}\right)^2.$$

(c) The first term does not depend on θ, and neither does the $\frac{1}{2\sigma^2}$ out front. Conclude that

$$\underset{\theta}{\operatorname{argmax}}\, LL(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right)^2.$$

Say out loud what you have just shown.

(d) Take the partial derivative to show

$$\frac{\partial}{\partial \theta_j}\left[-\sum_{i=1}^{n}\left(y^{(i)} - \theta^T x^{(i)}\right)^2\right] = \sum_{i=1}^{n} 2\left(y^{(i)} - \theta^T x^{(i)}\right) x_j^{(i)}.$$

Where exactly does the chain rule get used, and where does the minus sign go?

(e) Line this gradient up next to the two you already know:

logistic: $\sum_i \left[y^{(i)} - \sigma(\theta^T x^{(i)})\right] x_j^{(i)}$, neural net output layer: $[y - \hat{y}]\, h_j$, linear: $\sum_i 2\left[y^{(i)} - \hat{y}^{(i)}\right] x_j^{(i)}$.

What is the pattern? Write the update rule for gradient ascent on θ_j in one line.

Problem 5: Fitting a Line by Hand

A team's points scored, y, is modeled from a single feature x (points scored in the previous game, in tens). We have four training datapoints:

x	1	2	3	4
y	3	5	6	9

The model is $\hat{y} = \theta_0 + \theta_1 x$.

(a) Suppose we start with $\theta = [\theta_0, \theta_1] = [0, 2]$. Compute $\hat{y}^{(i)}$ and the error $y^{(i)} - \hat{y}^{(i)}$ for each of the four datapoints, and report the sum of squared errors.

(b) Compute both partial derivatives $\frac{\partial}{\partial \theta_0}$ and $\frac{\partial}{\partial \theta_1}$ of the negative sum of squared errors at this θ, using the formula from Problem 4(d). Remember that $x_0 = 1$ for every datapoint.

(c) Take one gradient ascent step with $\eta = 0.01$. Report the new θ, the new predictions, and the new sum of squared errors. Did it improve?

(d) The optimal fit for this dataset is $\theta_0 = 1.0$, $\theta_1 = 1.9$. Compute the four residuals $y^{(i)} - \hat{y}^{(i)}$ at the optimum, and verify that they sum to 0. Why must the residuals sum to zero at the optimum? *(Hint: look at the $j = 0$ component of the gradient.)*

(e) Report the sum of squared errors at the optimum, and compare it to your answers in (a) and (c).

Problem 6: The Noise Variance Is a Parameter Too

We quietly ignored σ^2 in Problem 4 — it dropped out of the argmax over θ. But it is a parameter of the model, and it is what lets us make *probabilistic* statements about our predictions rather than just point predictions.

(a) Using the optimal fit from Problem 5(d), the four residuals were $[0.1,\ 0.2,\ -0.7,\ 0.4]$. Using the Gaussian variance MLE $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \mu)^2$ with $\mu = 0$, estimate σ^2.

(b) A bike-share company fits a linear regression to predict hourly ridership. On a held-out test set of 5 datapoints the prediction errors are

$$[\,3,\ -5,\ 1,\ -4,\ 7\,].$$

Estimate σ^2 for the error random variable $Z \sim N(0, \sigma^2)$.

(c) Using your estimate, compute the probability that the model's prediction for a new hour is off by more than 10 riders in either direction.

(d) Why do we compute this on the *test* set rather than the training set? What would go wrong otherwise?

(e) Notice that this gives you a full predictive distribution, not just a number: $Y_{\text{new}} \mid X = x \sim N(\theta^T x, \hat{\sigma}^2)$. Write down a range that contains the prediction with probability about 0.95, for a new datapoint whose point prediction is 120 riders.

Problem 7: Reinforcement Learning

In supervised learning, every training datapoint comes with a label. In reinforcement learning, an agent takes actions in an environment and receives only a *reward*, possibly long after the action that earned it.

A critter lives forever. At each timestep $t = 0, 1, 2, \ldots$ it receives a reward R_t, and its objective is to maximize the expected **discounted return**

$$G = \sum_{t=0}^{\infty} \gamma^t R_t, \qquad 0 < \gamma < 1.$$

(a) Suppose $R_t = 1$ for every t, and $\gamma = 0.9$. Compute G. Which Lecture 5 fact did you use?

(b) Now suppose the R_t are i.i.d. with $E[R_t] = 0.5$, and $\gamma = 0.8$. Compute $E[G]$. Justify the step where you moved the expectation inside the sum.

(c) Explain in one or two sentences what would go wrong if we set $\gamma = 1$.

(d) The critter can eat red berries or green berries. Each berry gives reward 1 with unknown probability θ and reward 0 otherwise. Starting from a Beta$(1, 1)$ prior on each berry's θ, the critter has now eaten **7 rewarding red berries out of 10** and **2 rewarding green berries out of 2**. Write down the posterior distribution for each berry type and compute the posterior mean and standard deviation of each.

(e) A greedy critter always eats the berry with the higher posterior mean. Which does it choose? Give a concrete reason why that might be the wrong long-run decision, using your standard deviations from (d). What is this tension called?

(f) In one sentence each, explain why reinforcement learning is *not* just supervised learning with the reward as the label. Name at least two structural differences.

Challenge Problem (optional): What If the Noise Isn't Gaussian?

(a) Suppose we replace the Gaussian noise assumption with **Laplace** noise, whose PDF is $f(z) = \frac{1}{2b}e^{-|z|/b}$. Redo the derivation of Problem 4 and show that maximum likelihood now tells us to minimize $\sum_i \left|y^{(i)} - \theta^T x^{(i)}\right|$.

(b) A single datapoint in your training set has $y = 10{,}000$ when every other y is near 50. Which of the two objectives — squared error or absolute error — is more disturbed by it, and why? Relate your answer to how fast each PDF's tail decays.

(c) Now go the other way: instead of changing the noise, put a prior on the parameters. Assume $\theta_j \sim N(0, \tau^2)$ independently, and find $\theta_{\text{MAP}} = \text{argmax}_\theta \left[\log f(\theta) + \sum_i \log f(y^{(i)} \mid x^{(i)}, \theta)\right]$. Show that this is equivalent to minimizing

$$\sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 + \lambda \sum_j \theta_j^2$$

and identify λ in terms of σ^2 and τ^2. *(This is called ridge regression. You just derived it.)*

(d) What happens to λ as $\tau^2 \to \infty$? Interpret that in words.