CSCI 145 Problem 7: Prediction Is a Shadow

July 20, 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 7: Prediction Is a Shadow

Hold an object under a lamp and its shadow lands on the floor. The floor cannot reproduce the object, so it reproduces the closest thing it can. Least squares does the same to your labels: the column space of $\mathbf{X}$ is the floor, the label vector $\mathbf{y}$ is the object, and the fitted prediction $\hat{\mathbf{y}}$ is the shadow. This problem turns the picture into algebra: an exact formula for the shadow, and an accounting of how much of $\mathbf{y}$ it explains.

Throughout, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix with linearly independent columns (so $\mathbf{X}^\top \mathbf{X}$ is invertible), $\mathbf{y} \in \mathbb{R}^n$ is the label vector, $\mathbf{w}^\star = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$, and $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}^\star$. (Lecture's loss carried a $\frac{1}{n}$ out front; a positive constant does not move a minimizer, so we drop it here.)

Part A: The Normal Equations, Geometrically

Lecture proved that at the optimum the residual $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to *every* column of $\mathbf{X}$. Write it as the single equation $\mathbf{X}^\top \mathbf{r} = \mathbf{0}$, saying in a sentence why that matrix form packages d separate inner products. Then substitute $\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{w}^\star$ and conclude the **normal equations**:

$$\mathbf{X}^\top \mathbf{X}\mathbf{w}^\star = \mathbf{X}^\top \mathbf{y}, \qquad \text{so} \qquad \mathbf{w}^\star = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Now reconcile this with the other route to the same place: Problem 3 Part C computed $\nabla_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = 2\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$, so set that gradient to zero and confirm you get the identical equation. In one sentence, explain why "the gradient vanishes" and "the residual is perpendicular to the column space" have to be the same condition.

Part B: Prediction Is a Projection

Define the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \in \mathbb{R}^{n \times n},$$

so that Part A reads $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ for *every* label vector $\mathbf{y}$: one fixed matrix turns any labels into the closest point of $\text{col}(\mathbf{X})$, with no re-fitting. Show algebraically that $\mathbf{H}^2 = \mathbf{H}$; a matrix with this property is called a **projection**, since applying it twice does no more than applying it once. Then explain in one sentence why that has to be true *geometrically*: if $\hat{\mathbf{y}}$ is already the closest point of $\text{col}(\mathbf{X})$ to $\mathbf{y}$, what is the closest point to $\hat{\mathbf{y}}$ itself?

Part C: The Pythagorean Split, and R^2

Because $\mathbf{r}$ is perpendicular to $\text{col}(\mathbf{X})$ and $\hat{\mathbf{y}} \in \text{col}(\mathbf{X})$, the vectors $\hat{\mathbf{y}}$ and $\mathbf{r}$ are themselves perpendicular. Expand $\|\mathbf{y}\|_2^2 = \|\hat{\mathbf{y}} + \mathbf{r}\|_2^2$ as an inner product and cancel the cross term, exactly as lecture did for $\|\mathbf{y} - w\mathbf{x}\|_2^2$, to show

$$\|\mathbf{y}\|_2^2 = \|\hat{\mathbf{y}}\|_2^2 + \|\mathbf{r}\|_2^2.$$

This is the Pythagorean theorem, with $\mathbf{y}$ the hypotenuse and the prediction and residual as the two legs. Divide both sides by $\|\mathbf{y}\|_2^2$ to define

$$R^2 = \frac{\|\hat{\mathbf{y}}\|_2^2}{\|\mathbf{y}\|_2^2} = 1 - \frac{\|\mathbf{r}\|_2^2}{\|\mathbf{y}\|_2^2},$$

the fraction of $\mathbf{y}$'s squared length the model explains. (Statistics courses center $\mathbf{y}$ by its mean first; we skip that refinement here.) Argue from the split why R^2 always lies in $[0, 1]$, and say what a fitted model with $R^2 \approx 0$ looks like in the shadow picture. Then state the lesson of this problem in one sentence, starting with "Least squares".

Where this goes next. The formula you derived is exact, but next lecture asks what it *costs* and finds a cheaper iterative alternative for large d. And when $d > n$, the matrix $\mathbf{X}^\top\mathbf{X}$ is not invertible: infinitely many $\mathbf{w}$ fit exactly, so there is no unique shadow to find, and the Generalization problem at the end of the next unit tells you which one to pick.