CSCI 145 Problem 5: The Best Possible Forecast

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 5: The Best Possible Forecast

Even a perfect forecaster misses tomorrow's temperature. This problem asks how well the *best possible* forecaster does: one who knows the entire joint distribution of (X, Y), so that no finite dataset and no estimation error stand in the way, leaving only the fundamental limit of predicting a random Y from a correlated X. In lecture, we computed the answer for a 2×2 table of weather and ice-cream sales; here you prove the general theorem the table was hiding.

Throughout, X and Y are jointly distributed random variables, a predictor is any function f, and we score predictions by the expected squared error $\mathbb{E}[(Y - f(X))^2]$, where the expectation is over the joint distribution.

Part A: The Best Constant

Warm up with forecasters who ignore the evidence: $f(X) = c$ for a constant c. By adding and subtracting $\mathbb{E}[Y]$ inside the square, show that

$$\mathbb{E}[(Y - c)^2] = \mathrm{Var}(Y) + (\mathbb{E}[Y] - c)^2,$$

and conclude that the best constant is $c^* = \mathbb{E}[Y]$, with error exactly $\mathrm{Var}(Y)$. (Check against lecture: for the cloudy/sunny table, $c^* = 21$ and the error is 99.) Keep the proof technique in hand: Part B is the same add-and-subtract, one level up.

Part B: The Best Function

Now allow *any* predictor f, and let

$$f^*(x) = \mathbb{E}[Y \mid X = x]$$

be the regression function from lecture. By adding and subtracting $f^*(X)$ inside the square, show that

$$\mathbb{E}[(Y - f(X))^2] = \mathbb{E}[(Y - f^*(X))^2] + \mathbb{E}[(f^*(X) - f(X))^2] + 2\,\mathbb{E}\big[(Y - f^*(X))(f^*(X) - f(X))\big],$$

and then prove the cross term is zero. (Hint: compute its expectation in two stages, first over Y with X held fixed, then over X. With X held fixed, $f^*(X) - f(X)$ is just a number and can slide out of the inner expectation; what is $\mathbb{E}[Y - f^*(X) \mid X]$?) Conclude that for every predictor f,

$$\mathbb{E}[(Y - f(X))^2] \geq \mathbb{E}[(Y - f^*(X))^2],$$

with the excess error equal to $\mathbb{E}[(f^*(X) - f(X))^2]$: the price of a predictor is exactly its mean squared distance from the conditional mean.

Part C: The Floor

How wrong is the best forecaster? Show that the minimum risk from Part B is

$$\mathbb{E}[(Y - f^*(X))^2] = \mathbb{E}[\mathrm{Var}(Y \mid X)],$$

where $\mathrm{Var}(Y \mid X = x)$ is the variance of Y under the conditional distribution given $X = x$. Now put the three parts together on the lecture's cloudy/sunny table. A constant forecast is a predictor like any other, so Part B says the 99 of Part A must split into the floor you just derived plus the constant's mean squared distance from f^*; compute both pieces and check that they add to 99. Finally, state in one sentence what all of this says about a model that reports *zero* error on noisy test data.

Part D*: A Different Loss, a Different Forecast

The conditional *mean* won because we chose squared error. Suppose instead we score with absolute error, $\mathbb{E}[|Y - c| \mid X = x]$, for a constant forecast c at a fixed x. Let F be the conditional cumulative distribution function of Y given $X = x$, assumed continuous. Show that

$$\frac{\partial}{\partial c}\,\mathbb{E}\big[|Y - c| \,\big|\, X = x\big] = 2F(c) - 1,$$

(differentiating under the expectation is fine), and conclude the optimal forecast is the conditional *median*, where $F(c) = \frac{1}{2}$. So the "best possible forecast" depends on how we price a miss, which is the observation next lecture builds on when it asks where a loss function comes from in the first place.

Where this goes next. The floor $\mathbb{E}[\mathrm{Var}(Y \mid X)]$ you derived in Part C reappears *verbatim* as the noise term when we decompose test error in the Methodology lecture, and the "condition on X, then average" move from Part B is the workhorse behind the variance-reduction proofs in the reinforcement learning unit.