Predictions and Decision Theory

36-462/662, Spring 2022

27 January 2022 (Lecture 4)

Housekeeping

- Back on campus starting next week
- Homework 0: tonight at 6 pm via Canvas
- Homework 1: tonight at 6 pm via Gradescope
- Homework 2: releasing tomorrow morning (if not before)

Previously

- We've looked at linear regression, linear classifiers and logistic regression as predictive methods
- In general: We want to use data to learn rules which we can be confident will predict well on average on new cases
- All the terms in that phrase have to be made precise
- Today we're going to focus on "rules" and "predict well on average"

Prediction

- Prediction is a guess about some event we haven't seen yet, but could see
  - Inference, but to an observable, not a parameter of the distribution
  - "The next roll of these 3 dice will be 18" vs. "The variance of rolling 3d6 is 8.75"
- We're interested in predictions done according to *rules*
- Rules are functions from inputs to outputs
  - We don't need to presume the *actual* target is a function of the inputs

Good and bad predictions

- We need a way of saying whether a rule is working well or not
- Predictions that come true are better than those that don't
- But are all mistakes equally bad?
  - Predicting 6 inches of snow when the reality is 5 seems better than predicting 10 inches, or 0 inches
  - Predicting someone's healthy when they're sick seems worse than the other way around
- This is where decision theory comes in

The elements of a decision problem

1. Possible **actions** A
2. **Information** X, which we get to see before taking an action
3. **States** Y picked by Nature
4. A **strategy** s is a function from X (information) to A (action)

- There is usually some class of strategies S available

5. A **loss function** $\ell(y, a)$: how much it hurts to take action a when the state is y

- The loss function is crucial but not enough on its own

The risk of a strategy

- The **risk** of a strategy is its expected loss, averaging over X and Y

$$r(s) = \mathbb{E}\left[\ell(Y, s(X))\right]$$

- This assumes that X and Y are both random variables with a joint distribution, say $P(X, Y)$
  - For now, our actions and strategy don't change P
  - We'll come back to decisions where our actions matter later in the course

Risk minimization

- Loss is bad, risk is expected loss $\Rightarrow$ try to minimize risk
- Use the law of total expectations:

$$\mathbb{E}\left[\ell(Y, s(X))\right] = \mathbb{E}\left[\mathbb{E}\left[\ell(Y, s(X))|X\right]\right]$$

  - Inner expectation is the **conditional risk**
- Now define

$$\sigma(x) \equiv \operatorname*{argmin}_{a \in A} \mathbb{E}\left[\ell(Y, a)|X = x\right]$$

  - Take the action that minimizes the conditional expected loss
  - "Do what's best, given what you know"

Minimizing the conditional risk really is optimal

- Minimizing the conditional risk everywhere minimizes the over-all risk:

$$\sigma = \operatorname*{argmin}_{s: X \mapsto A} \mathbb{E}\left[\ell(Y, s(X))\right]$$

- This is worth proving
- It's enough to show that for any other strategy s, $r(s) - r(\sigma) \geq 0$ (why?)

$$\begin{aligned} r(s) - r(\sigma) &= \mathbb{E}\left[\ell(Y, s(X)) - \ell(Y, \sigma(X))\right] & (1) \\ &= \mathbb{E}\left[\mathbb{E}\left[\ell(Y, s(X)) - \ell(Y, \sigma(X))|X\right]\right] & (2) \end{aligned}$$

- So for each x,

$$\mathbb{E}\left[\ell(Y, s(x))|X = x\right] \geq \mathbb{E}\left[\ell(Y, \sigma(x))|X = x\right]$$

- Write r_0 for the minimal risk $r(\sigma)$
  - Generally not 0 (as we've seen with regression and classification)

Minimizing the risk in a class of strategies

- Remember S is the strategies we can actually use
- Typically doesn't contain σ so we do the best we can:

$$s^* = \operatorname*{argmin}_{s \in S} r(s)$$

- $r(s^*) \geq r_0$, maybe much larger, maybe only a little

The approximation-estimation trade-off

- A basic decomposition: for any strategy s,

$$r(s) = r_0 + (r(s^*) - r_0) + (r(s) - r(s^*))$$

- r_0 = true minimum risk
- $r(s^*) - r_0$ = **approximation error** (due to using S)
- $r(s) - r(s^*)$ = **estimation error** (due to not using s^*)
- Generally:
  - Making S larger reduces approximation error (better optimum)
  - Making S larger increases estimation error (harder to find the optimum)
- We will come back to this over and over through the course

Back to prediction problems

1. Actions = predictions
2. Information = covariates, regressors, features (etc.)
3. States = the target variable we're trying to predict
4. Strategy = prediction rule = function from information to actions
5. Loss function = ?

- Different loss functions will give us different risks for the same strategy
- Different loss functions will lead to different optimal prediction rules

Regression, for example

1. Actions = predictions = real numbers = guesses at the regressand
2. Information = vectors of real numbers = covariates, regressors ("independent variables")
3. States = "dependent variable", "regressand"
4. Strategy = prediction rule = regression function
5. Loss function = ?

- The usual loss function is squared error, $\ell(y, a) = (a - y)^2$
- Risk then is **expected squared error**
- The minimizer of $\mathbb{E}\left[(Y - a)^2\right]$ is $a = \mathbb{E}[Y]$
- The minimizer of $\mathbb{E}\left[(Y - a)^2 | X = x\right]$ is $a = \mathbb{E}[Y|X = x]$
- The true or optimal regression function is $\mu(x) = \mathbb{E}[Y|X = x]$, the conditional mean function

Linear regression, for example

- Generally the conditional mean function is very nonlinear in x
- What if we're only allowed to use linear functions of x?
- We know the answer to this one:

$$s^*(x) \quad = \quad \mathbb{E}[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]}(x - \mathbb{E}[X]) \tag{3}$$

  The expected squared error is

$$\mathbb{E}\left[(Y - s^*(X))^2\right] = \text{Var}[Y] - \frac{(\text{Cov}[X, Y])^2}{\text{Var}[X]} = r(s^*)$$

- (Similarly for multivariate X but more linear algebra)

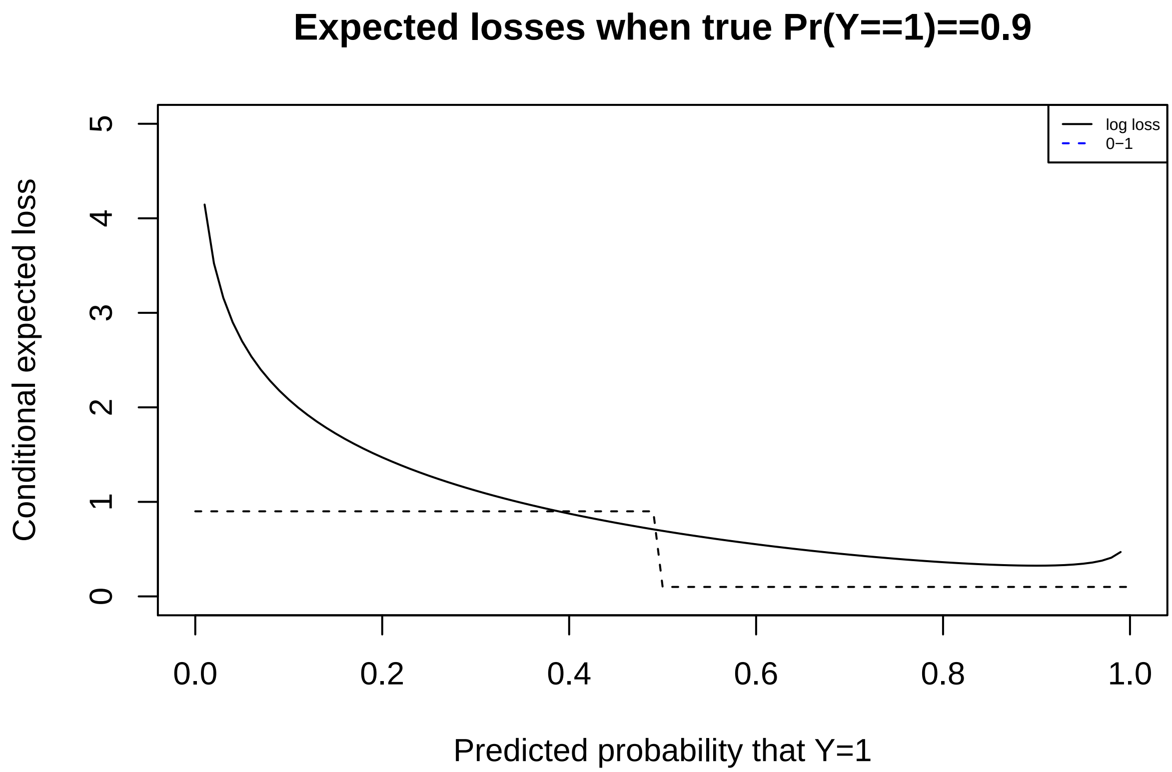
Alternative loss functions for regression

- Remember all this is with squared error as the loss function
- **Absolute error**, $\ell(y, a) = |y - a|$
  - Risk minimized with median, not mean
- **0-1 or Hamming error**: 0 if $y = a$, 1 if $y \neq a$
  - Risk minimized with the mode
- **Huber's robust error**, continuously switch over from absolute error to squared error
  - No closed form for the optimal action
- **Tolerance region**: zero error if $|y - a| \leq \epsilon$, then growing (say) linearly in $|y - a|$
  - Also no closed form
- Asymmetric errors if over-shooting is better (or worse) than under-shooting
- Some of these are easier to work with than others, but that doesn't make them application-appropriate

Some losses for classification

- Classification = predicting a categorical variable
- **0-1 loss**: $\ell(y, a) = 0$ if $a = y$, $\ell(y, a) = 1$ if $y \neq a$
  - Makes sense when the actions are class labels
  - Minimized by predicting the most probable class
- **Weighted losses**: $\ell(y, a) = L_{ya}$ for some matrix, says how bad it is to predict a when the reality is y
  - e.g. "you said this person didn't have cancer when they really did" vs. "you made this person go in for additional tests when they were fine"
  - also makes sense when the actions are class labels
- Maybe we predict the probability that $Y = 1$ (rather than $Y = 0$) so $A = [0, 1]$
- **Log loss**: $\ell(y, a) = -y \log a - (1 - y) \log (1 - a)$

0-1 loss vs. log loss



- 0-1 loss just cares if your probability is on the correct side of 1/2
- Log loss wants you to get the probability just right, gets more upset when you're confident *and wrong*
- Smooth functions (like log loss) are often easier to work with theoretically and computationally, but 0-1 is more forgiving of getting the distribution wrong…
- Choosing a loss function is not something decision theory helps us with…

Other possible loss functions

- "How long did the user stay on our site"?
- "Did the user click on an ad?"
- "How much money did we make from this transaction?"
- "Did the patient live?"
- "How much did treating this patient cost us?" -(Some of these are good things, so "loss" $= -$ good thing)

Connecting to data

- I promised we'd focus on the "rules" and "predict well on average" parts of "learn rules from data that will predict well, on average, on new cases"
- Rules are strategies
- "predict well on average" = low risk
- Risk is *defined* as an expectation using the true distribution, $\mathbb{E}\left[\ell(Y, s(X))\right] = \int \ell(y, s(x))p(x, y)dxdy$
- We don't know the true distribution $p(x, y)$
- We just have limited data
- How can we minimize risk?

Connecting to data

- Natural idea: minimize the average risk *on the data*

$$\widehat{r}_n(s) \equiv \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, s(x_i))$$

  - Often called the **empirical risk**
- By law of large numbers, $\widehat{r}_n(s) \rightarrow r(s)$ as $n \rightarrow \infty$, for any *fixed* s
- **Empirical risk minimization**: Pick the rule/strategy that minimizes the empirical risk

$$\hat{s} \equiv \underset{s \in S}{\operatorname{argmin}}\, \widehat{r}_n(s)$$

  - "Pick the rule that did best, on average, on the data you have"
  - Least squares and maximum likelihood are both examples of ERM
- To understand when this works, how it works,. and what else we might do, we're going to have to know understand a bit more about optimization...

Back-up: Alternatives to minimizing risk

- Risk is expected loss
- Other things we could minimize:
  - Median loss
  - 95th (99th, 99.9999th) percentile of loss ($\approx$ "value at risk" in fiance)
  - Maximum loss (**minimax**)
  - Probability of one specific type of error (false negative, false positive)
- We could not minimize at all:
  - Any strategy with a risk (median loss, etc.) below some threshold is OK ("satisficing" instead of optimizing)
  - Any strategy where $\mathbb{P}(\ell(Y, s(X)) > \epsilon) < \delta$ is OK
- But risk is traditional:
  - It makes sense if you're working "actuarially", looking for rules that will be OK applied across a large population
  - Minimax can get pretty paranoid (what if the Moon is really an alien trap?)
  - The math is clean
  - Preferences that meet some axioms can be "rationalized" as minimizing risk nn * Some of the axioms are hard to swallow
  - There's a lot of tradition to draw on

Back-up: Why decision theory?

- Jerzy Neyman (2nd greatest statistician of the 20th century): forget about inductive *inference*, study rules of inductive *behavior*
- Abraham Wald: reformulates inference as decision problems, shows how to connect to practical things like quality control and how to fight WWII
- Statistical theorists everywhere after the war: yes! use decision theory to find optimal procedures for all the inference problems!
- Statistical learning: inherited decision theory from theoretical statistics
  - The people coming from computer science were, at least to begin with, fixated on what we'd call 0-1 loss for classification, and situations where the minimum risk was exactly 0

Back-up: Loss vs. utility, "risk" vs. "risk"

- Statisticians like to work with loss functions, and minimize expected loss
- Economists like to work with utility functions, and maximize expected utility
- Insert a minus sign to turn one into the other
- In business and finance, they like to maximize returns in dollars (or yuan, etc.)
  - Economists would say that the utility of each extra unit of money is declining, so maximizing expected profit is not necessarily maximizing expected utility
    - And taxing the rich at higher rates than the poor is straightforwardly better, in terms of utility, than a flat tax…
- In business and finance, "risk" is (basically) defined as the *variance* of the monetary returns
  - Occasionally leads to confusion