Machine learning

Risk: probability and loss

Joshua Loftus

Randomness

- Minimizing squared error *on observed data*

$$\text{minimize } \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Plug-in principle: assuming a probability model, i.e. some joint distribution $p_{X,Y}(x, y)$

$$\text{minimize } \mathbb{E}[(Y - \alpha - \beta X)^2]$$

Generative ML

- Some machine learning methods do not explicitly use probability distributions
- Those that do use probability are sometimes called "generative models" because they
  1. Model the "data generation process" (DGP)
  2. Can be used to generate (synthetic) "data" (sampling with a random number generator)

This course is mainly focused on methods that do use probability, and we will always try to do so explicitly/transparently (not hiding our assumptions)

Conditional distributions

Within generative machine learning, supervised learning is broadly about modeling the *conditional distribution of the outcome given the features*

$$p_{Y|X}(y|x) = p_{X,Y}(x,y)/p_X(x)$$

Some methods try to learn this entire distribution, others focus on some summary/functional, e.g.

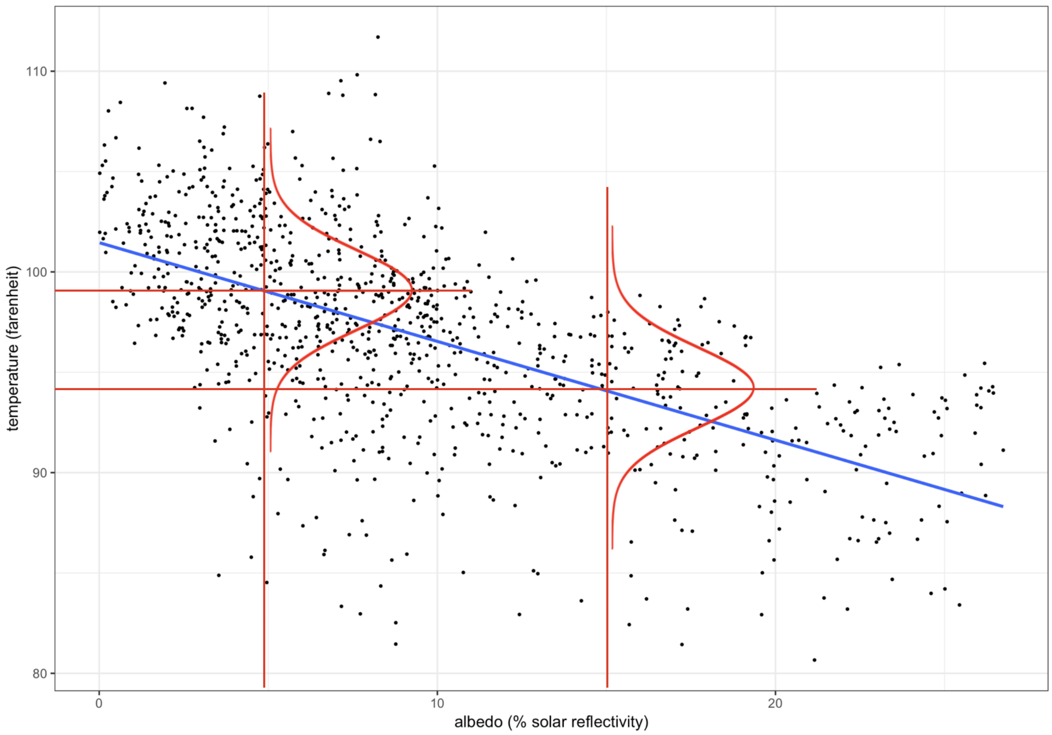
conditional expectation

$$\mathbb{E}_{Y|X}[Y|X]$$

or conditional quantile

$$Q_{Y|X}(\tau)$$

(for the τth quantile)



Curves showing $p_{Y|X}(y|x)$ at two values of x

A variety of objectives

It can be shown (another good **Exercise**!) that

- The conditional expectation function (CEF)

$$f(x) = \mathbb{E}_{Y|X}[Y|X = x]$$

minimizes the expected squared loss

$$f(x) = \arg\min_g \mathbb{E}_{X,Y}\{[Y - g(X)]^2\}$$

- Similarly, **quantile regression** is about, e.g.

$$Q_{Y|X}(0.5) = \arg\min_g \mathbb{E}_{X,Y}[|Y - g(X)|]$$

(for other quantiles, "tilt" the absolute value loss function)

Risk = expected loss

Other examples also fit into this broad framework

For a given **loss function** $L(x, y, g)$, find the optimal "regression" function $f(x)$ that minimizes the risk, i.e.

$$R(g) = \mathbb{E}_{X,Y}[L(X, Y, g)]$$

$$f(x) = \arg\min_g R(g)$$

Statistical machine learning:

$$\mathbb{E} \longleftrightarrow \frac{1}{n}\sum$$

Algorithms can leverage LLN, CLT, subsampling, etc...

Our focus

- For now, squared error. Other cases similar! (Bias-variance)
- Later: categorical outcome loss functions (classification)

Additional modeling assumptions

Linear regression is based on an *assumption* that the conditional expectation function (CEF) is (*or can be adequately approximated as*) linear

$$f(x) := \mathbb{E}_{Y|X}(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

(**Question**: why no ε errors in this equation?)

Statistical wisdom

Sometimes this assumption works marvelously

Other times it breaks spectacularly

Often, it's somewhere in the gray area

"All models are wrong, but some are useful"

Always, always, *always* remember George Box:

> Since all models are wrong *the scientist must be alert to what is* ***importantly wrong***. It is inappropriate to be concerned about mice when there are tigers abroad.

Strengths of machine learning

- Relaxing the linearity assumption and using flexible, non-linear models
- Specialized methods for high-dimensional linear regression, where there are many predictor variables, possibly even $p > n$
- Beating other approaches at pure prediction accuracy, trading off simplicity/interpretability for better predictions

Recently, people have started caring more about interpretability again -- an emphasis in this course