Distributions as Target Variables for Deep Learning Model

University of Washington Department of Atmospheric and Climate Science Machine Learning Journal Club

9/18/2025

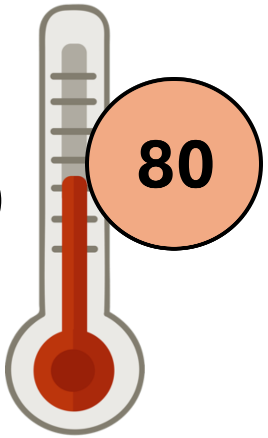
Miles Epstein

Coding Workshop

https://github.com/milesepstein13/ml-distribution-workshop

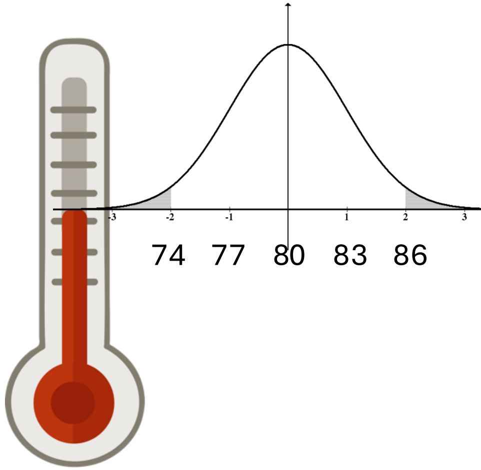
Basic Idea

- Scalar Regression Model:
  - Predicts *scalar* as target variable
  - Many possible architectures (linear regression, ANN, CNN, etc.)



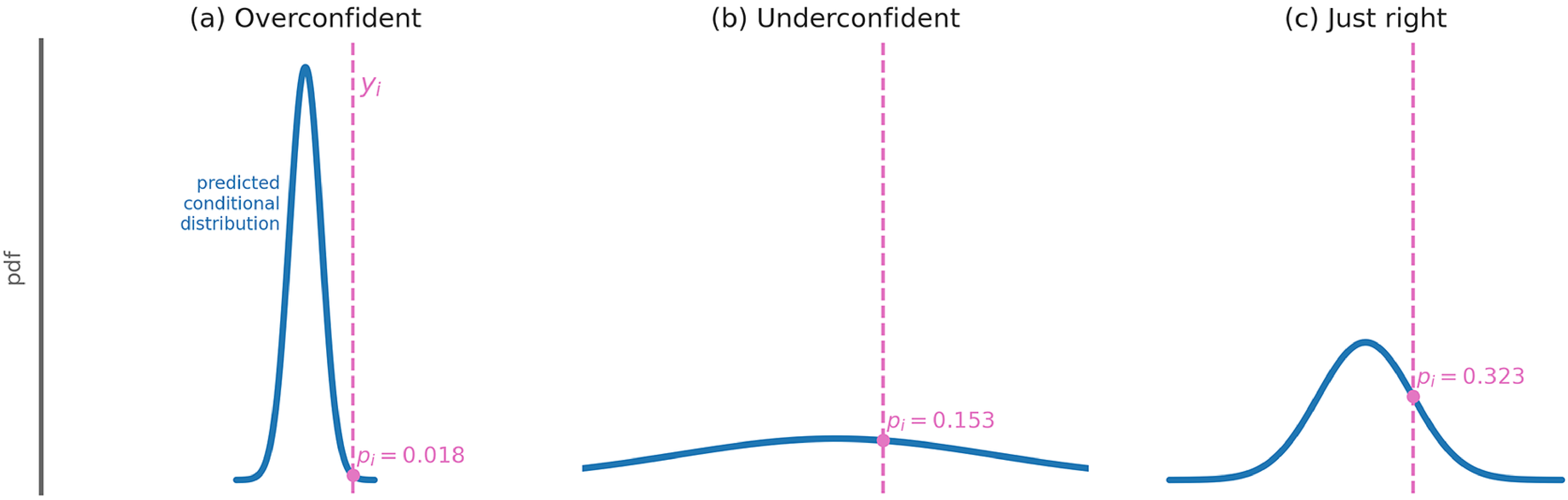
e.g. “I predict the high temperature will be 80 °F tomorrow”

- Distributional Regression Model
  - Predicts *multiple scalars* that are *parameters* of a probabilistic distribution for the target variable



e.g. “I predict the high temperature tomorrow will come from a Gaussian with mean 80 and SD 3”

Basic Idea



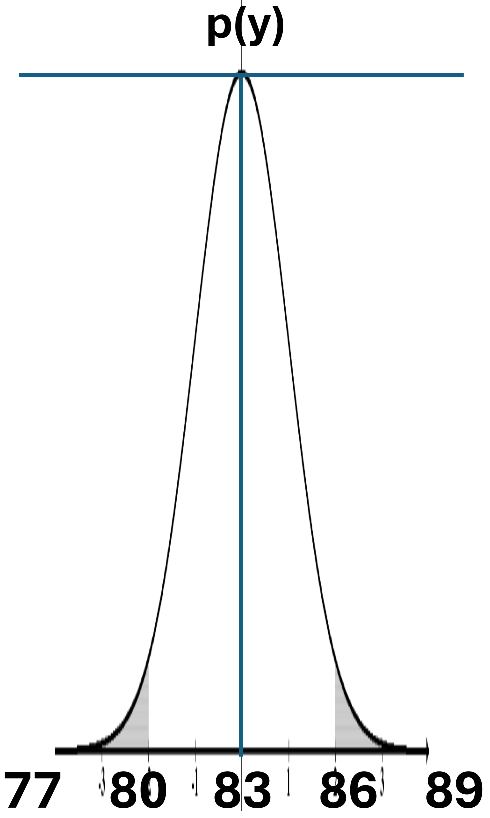
- Figure From Barnes et al. (2023)

In Literature

- Concept introduced with Gaussian for dummy dataset:
  - Estimating the Mean and Variance of the Target Probability Distribution (Nix and Weigend 1994)
- Use of non-Gaussian distributions:
  - Generalized Additive Models for Location, Scale and Shape (Rigby and Stasinopoulos 2005)
- Widespread use in Computer Science, starting to see use in Atmospheric Science:
  - Controlled Abstention Neural Networks for Identifying Skillful Predictions for Regression Problems (Barnes and Barnes 2021)
  - Sinh-arcsinh-normal distributions to add uncertainty to neural network regression tasks: Applications to tropical cyclone intensity forecasts (Barnes et al. 2023): Uses SHASH distribution instead of Gaussian

Loss Function

- Consider: Observed temperature y = 83
- Scalar Regression (prediction $\hat{y} = 80$):
  - (R)MSE: $(83 - 80)^2$ = 9
  - MAE: $|83 - 80| = 3$
- Distributional Regression (prediction $\hat{y} = \mathcal{N}(80, 3)$)
  - What represents the quality of prediction?
    - Likelihood: $L(y|x) = p(y|x; \theta(x))$
    - Take negative log for easy optimization
    - **Loss** $\ell = \sum_{i=1}^{N} -\boldsymbol{p}(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\theta}(\boldsymbol{x}_i))$
    - ***Negative Log Likelihood***



Our case:

Our case: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

- Recall:
- $L(y|x) = \frac{1}{1.5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{83-83}{1.5}\right)^2}$
- $\ell = -\ln\left(\frac{1}{1.5\sqrt{2\pi}} e^{0}\right)$
- $\boldsymbol{\ell = -\ln\left(\frac{1}{1.5\sqrt{2\pi}}\right) = -.51}$

More generally, for gaussian:

$$-\ln p(y \mid \mu, \sigma^2) = -\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big)\right]$$

$$= -\Big(-\tfrac{1}{2}\ln(2\pi\sigma^2) - \tfrac{(y-\mu)^2}{2\sigma^2}\Big)$$

$$= \tfrac{1}{2}\ln(2\pi\sigma^2) + \frac{(y-\mu)^2}{2\sigma^2}$$

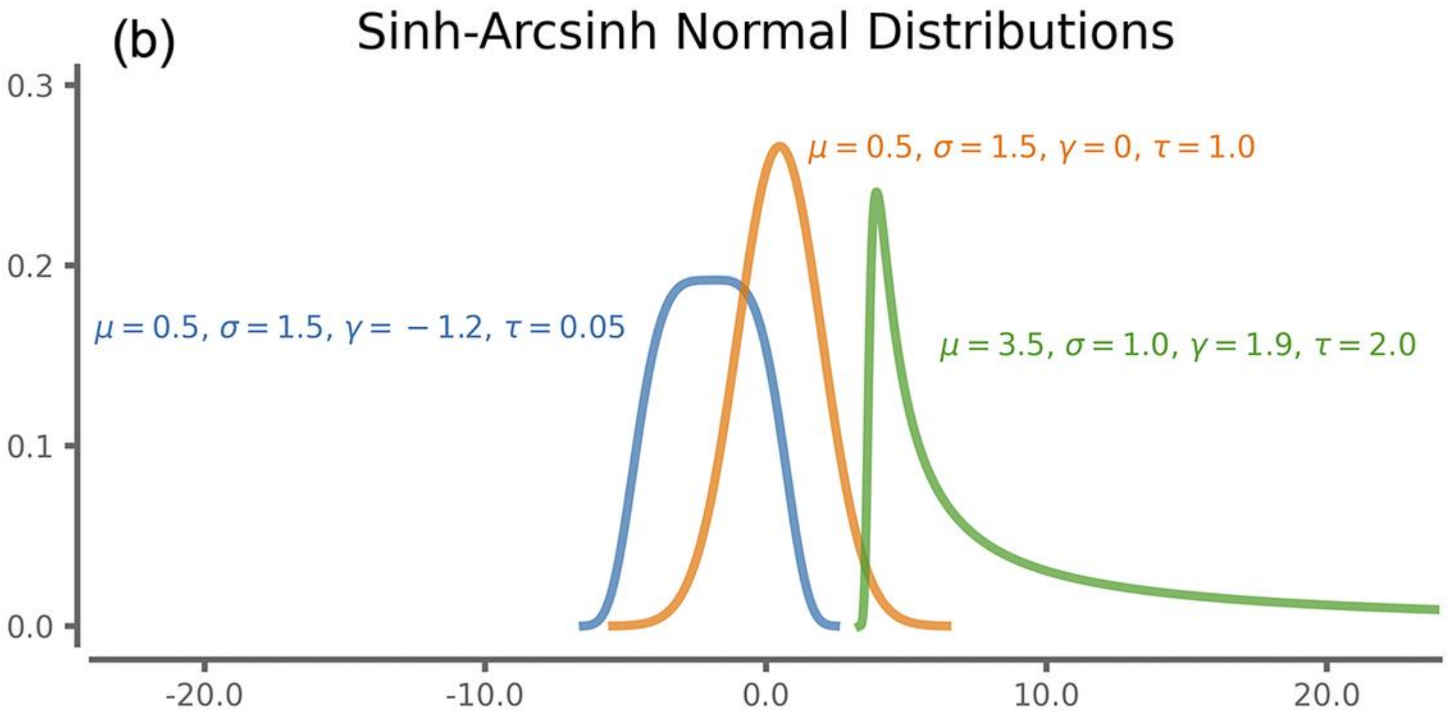
$$\propto \quad \frac{1}{2}\ln\sigma^2 + \frac{(y-\mu)^2}{2\sigma^2}$$

SHASH

$$f(x) = \frac{\delta}{\eta} \cdot \sqrt{\frac{1 + S^2(y;\epsilon,\delta)}{2\pi(1+y^2)}} \cdot \exp\left(-\frac{1}{2}S^2(y;\epsilon,\delta)\right),$$

where

$$y = \frac{x - \xi}{\eta}. \qquad S(\chi;\alpha,\beta) = \sinh(\beta \cdot \text{asinh}(\chi) - \alpha).$$



- Four parameters:
  - Location (ξ/μ; analogous to mean)
  - Scale (η/σ; analogous to variance)
  - Skewness (ε/γ; asymmetry, positive skews right)
  - Tailweight (δ/τ; larger means heavier tails and more "peaked")
- As with Gaussian,
  - Model predicts these four parameters
  - Loss is NLL (from PDF above)
- Don't worry, I won't make you do any of the math

Discussion

- What are some advantages/disadvantages of distributional regression?
- What situations could we apply this method in?

Coding Workshop

https://github.com/milesepstein13/ml-distribution-workshop

Fall Meeting Schedule

https://www.when2meet.com/?32421861-wWdvB

or

https://tinyurl.com/3s24n4wr