ST790 — Fall 2022
Imprecise-Probabilistic Foundations of Statistics

Ryan Martin
North Carolina State University
`www4.stat.ncsu.edu/~rmartin`

Week 13a

This lecture

- Regression problems in ML
- Imprecise-probabilistic approaches:
  - IMs and conformal prediction
  - Denoeux's[1] *random fuzzy sets*
- ...

[1] See, also, Cuoso & Sanchez (*Fuzzy Sets & Systems* 2011)

Quick recap of ML

- Ingredients:
  - data, e.g., features X_i and labels/responses Y_i
  - class $\mathcal{F}$ of functions, hopefully $y \approx f^\star(x)$ for some $f^\star \in \mathcal{F}$
  - loss function, ℓ_f, to rate quality of f
- Note the absence of a statistical model...
- Training step boils down to "estimating" f by minimizing an empirical risk function, i.e.,

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_f(X_i, Y_i)$$

- Might need penalty terms to manage complexity, numerical methods (e.g., SGD) will probably be needed too
- Use the trained $\hat{f}_n$ to predict/classify new examples

Regression case

- Now focus on *continuous* responses Y_i
- Common choice of loss: $\ell_f(x, y) = \{y - f(x)\}^2$
- Familiar case of *linear model*
  - $\mathcal{F} = \{x \mapsto \beta^\top x : \beta \in \mathbb{R}^q\}$
  - then $\hat{f}_n$ is the least-squares fitted mean response
- "Linear models are too restrictive" — *not true!*
- In fact, linear models are basically all we know:
  - $\mathcal{F} = \{x \mapsto \beta^\top B^{(m)}(x) : \beta \in \mathbb{R}^m\}$ for fixed m
  - ...
  - neural networks, etc., are basically (high-dim) linear models in transformed x's

Regression case, cont.

- Relationship between y and x could be complex:
  - x itself is high-dim
  - flexibility baked into $\mathcal{F}$ introduces high-dim parameter
- In that case, the empirical risk minimization takes the form

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_f(X_i, Y_i) + \lambda \, \mathrm{pen}(f) \right\}$$

- New pieces:
  - $\mathrm{pen}(f)$ is a penalty, larger when f is more complex
  - e.g., smooth functions are less complex, and sparse β typically makes smoother f_β, so $\mathrm{pen}(f_\beta) = \|\beta\|_1$ is reasonable
  - λ is weight to balance influence of penalty
- Then the solution, $\hat{f}_n$, depends on (λ, pen)

Uncertainty quantification

- Given $\hat{f}_n$, we only know how to predict Y_{n+1}
- Again, no matter how fancy the model/algorithm, there's no guarantee that $\hat{f}_n(X_{n+1})$ exactly equals Y_{n+1}
- How to quantify uncertainty?
- With simple models and/or strong assumptions, this is relatively easy to do, solutions would more-or-less agree
- More generally: this is challenging/non-trivial
- Imprecise-probabilistic ideas:
  - Cella and M., `arXiv:2112.10234`
  - Denoeux, `arXiv:2202.08081` and `arXiv:2208.00647`

Conformal prediction and valid IMs

- Conformal prediction algorithm...
- The output $\pi_n(y \mid x)$ defines a x-dependent possibility contour on $\mathbb{Y}$, leads to a possible measure with

$$\overline{\Pi}_n(A \mid x) = \sup_{y \in A} \pi_n(y \mid x), \quad A \subseteq \mathbb{Y}$$

- An IM[2] that achieves *strong prediction validity*, i.e.,

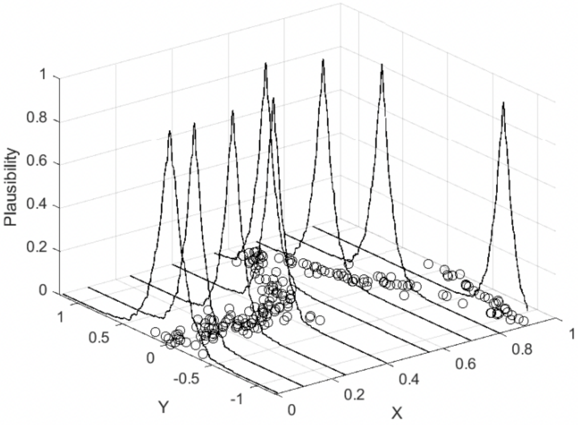
$$\sup_{\text{exchangeable P}} \mathsf{P}\{\pi_n(Y_{n+1} \mid X_{n+1}) \leq \alpha\} \leq \alpha, \quad \alpha \in [0,1]$$

- Doesn't require "correctly specified models"[3]
- Predictive probability distributions can't be valid in this sense

[2] Connection between IMs and CP is made on the random set level; this *get-CP-first-then-interpret-as-an-IM* is simpler to explain

[3] More efficient when fitted model is "right"—see, e.g., Slide 20 below

Conformal prediction and valid IMs, cont.



Random fuzzy numbers

- New extension of DS theory & possibility theory
- Basic road-map:
  - neural net provides a flexible relationship between Y and X
  - observations (X_i, Y_i) allow for learning this relationship
  - quantify uncertainty about Y_{n+1}, given X_{n+1} and training data, via a suitable random fuzzy number
- Need some more background:
  - fuzzy numbers
  - random version thereof

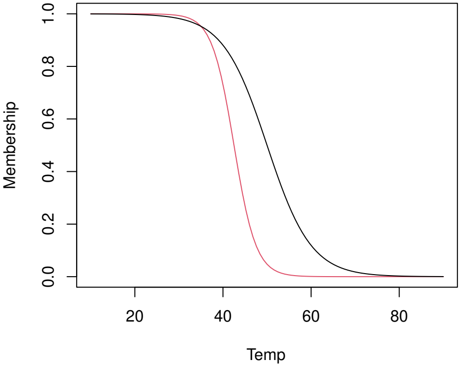
Fuzzy sets

- Fuzzy ___ deals with ambiguity (Zadeh 1960s)
  - Fuzzy logic generalizes Boolean true-or-false logic
  - e.g., "Sam is tall" isn't universally true or false
- Fuzzy sets generalize the notion of a (crisp) set
- Fuzzy sets[4] $\tilde{A}$ in a space $\mathbb{Y}$ are determined by a *membership function*, say, $\mu_{\tilde{A}} : \mathbb{Y} \to [0, 1]$
  - a fuzzy set is *crisp* if $\mu_{\tilde{A}}(y) \in \{0, 1\}$ for all y
  - membership function of a crisp set is its indicator function
- Membership function assigns to each $y \in \mathbb{Y}$ a quantitative *degree of membership*, $\mu_{\tilde{A}}(y) \in [0, 1]$, to the fuzzy set $\tilde{A}$
- A fuzzy set is characterized by its membership function

[4] Customary to write A for an ordinary set and $\tilde{A}$ for a fuzzy version

Fuzzy sets, cont.

- $\tilde{A} = \{\text{cold temperatures when I lived in Chicago}\}$
- $\tilde{B} = \{\text{cold temperatures after I moved to NC}\}$



Fuzzy sets, cont.

- Just like for crisp sets, there's a *fuzzy set arithmetic*[5]
- All amount to operations with membership functions
- I'll focus just on *fuzzy set intersection*[6]
- For fuzzy sets $\tilde{A}$ and $\tilde{B}$, the intersection $\tilde{A} \cap \tilde{B}$ is defined by the membership function

$$\mu_{\tilde{A}\cap\tilde{B}}(y) = \mu_{\tilde{A}}(y) \star \mu_{\tilde{B}}(y)$$

- "$\star$" is a *t-norm*,[7], e.g., $a \star b = ab$ or $a \star b = a \wedge b$

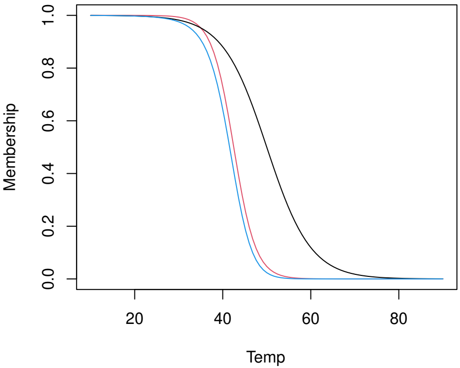
[5] e.g., see Hanss's *Applied Fuzzy Arithmetic*

[6] a.k.a. "conjunctive combination"

[7] Triangular norm: `https://en.wikipedia.org/wiki/T-norm`

Fuzzy sets, cont.

- $\tilde{A} = \{\text{cold temperatures when I lived in Chicago}\}$
- $\tilde{B} = \{\text{cold temperatures after I moved to NC}\}$
- $\tilde{A} \cap \tilde{B}$, product t-norm



Fuzzy sets, cont.

- Close connection to possibility theory
- The *height* of $\tilde{A}$ is $\sup_y \mu_{\tilde{A}}(y)$
- If height equals 1, then
  - $\tilde{A} \longleftrightarrow$ possibility measure
  - former's membership fn is the latter's contour fn
- Like Dempster's rule combines of belief functions, the fuzzy set intersection can combine possibility measures
- That is, if π_1 and π_2 are possibility contours on $\mathbb{Y}$, then these can be combined to a new possibility contour[8] as

$$\pi_{1\star 2}(y) = \frac{\pi_1(y) \star \pi_2(y)}{\sup_v\{\pi_1(v) \star \pi_2(v)\}}, \quad y \in \mathbb{Y}$$

[8] See Week 09b, Slide 14

Random fuzzy sets

- On a prob space $(\Omega, \cdot, \mathsf{P})$, define $\tilde{Y} : \Omega \to [0,1]^{\mathbb{Y}}$ s.t.

$$\tilde{Y}(\omega) \quad \text{is a fuzzy set in } \mathbb{Y} \text{ for each } \omega \in \Omega$$

- Defines a *random fuzzy set*
- If $\text{height}(\tilde{Y}) = 1$ P-a.s., then there's a *random possibility meas*

$$\text{poss}_{\tilde{Y}}(A) := \sup_{y \in A} \mu_{\tilde{Y}}(y), \quad A \subseteq \mathbb{Y}$$

- The UQ on $\mathbb{Y}$ provided by a random fuzzy number can be described by the upper probability

$$\overline{\Pi}(A) = \int \text{poss}_{\tilde{Y}(\omega)}(A) \, \mathsf{P}(d\omega), \quad A \subseteq \mathbb{Y}$$

- Corresponding lower probability, $\underline{\Pi}$, is a belief function

Random fuzzy sets, cont.

- For one or more random fuzzy sets (RFSs), the evidence they contain can be pooled using fuzzy set intersection
- The rule is associative, so it suffices to explain for two RFSs
- Roughly:
  - two indep pieces of evidence $(\Omega_j, \cdot, \mathsf{P}_j, \tilde{Y}_j)$, $j = 1, 2$
  - $(\tilde{Y}_1 \cap \tilde{Y}_2)(\omega_1, \omega_2)$ has membership function

$$\mu_{(\tilde{Y}_1 \cap \tilde{Y}_2)(\omega_1,\omega_2)}(y) \propto \mu_{\tilde{Y}_1(\omega_1)}(y) \star \mu_{\tilde{Y}_2(\omega_2)}(y), \quad y \in \mathbb{Y}$$

  - final UQ obtained by averaging wrt $\mathsf{P}_1 \times \mathsf{P}_2$
- With n pieces of evidence/RFSs, final UQ on $\mathbb{Y}$ is

$$\overline{\Pi}(A) = \int \mathrm{poss}_{\tilde{Y}_1 \cap \cdots \cap \tilde{Y}_n}(A)\, d(\mathsf{P}_1 \times \cdots \mathsf{P}_n)$$

Gaussian random fuzzy numbers

- A *Gaussian fuzzy number*[9] corresponds to a membership fn

$$\mu(y) = \exp\{-\tfrac{h}{2}(y-m)^2\}, \quad y \in \mathbb{R}$$

- Parametrized by a mean m and precision $h \geq 0$
- Key property: Gaussian fuzzy numbers are closed under fuzzy set intersection
- A *Gaussian random fuzzy number* $\tilde{Y}$ is a Gaussian fuzzy number with a mean $M \sim \mathrm{N}(\theta, \sigma^2)$, i.e.,

$$\mu_{\tilde{Y}(\omega)}(y) = \exp[-\tfrac{h}{2}\{y - M(\omega)\}^2], \quad y \in \mathbb{R}, \quad \omega \in \Omega$$

- Notation: $\tilde{Y} \sim \tilde{\mathrm{N}}(\theta, \sigma^2, h)$

[9] A "fuzzy number" is just a fuzzy interval

Evidential neural net regression

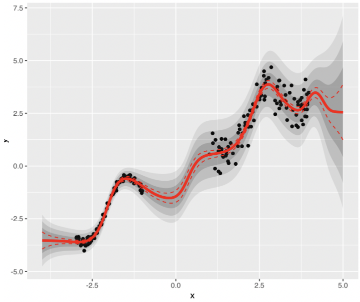
- Prototypes $w_1, \ldots, w_J$ in $\mathbb{R}^q$
- For a generic $x \in \mathbb{R}^q$ and for prototype $j = 1, \ldots, J$:
  - activation of prototype j, $a_j(x) = \exp\{-\gamma_j^2 \|x - w_j\|^2\}$
  - mean function $\mu_j(x) = \alpha_j + \beta_j^\top x$
  - GRFN $\tilde{Y}_j(x) \sim \tilde{\mathrm{N}}(\mu_j(x), \sigma_j^2, a_j(x) h_j)$
- Take fuzzy set intersection of the j-specific GRFNs...
- Gives $\tilde{Y}(x) \sim \tilde{\mathrm{N}}\big(\mu(x), \sigma^2(x), h(x)\big)$, with, e.g.,

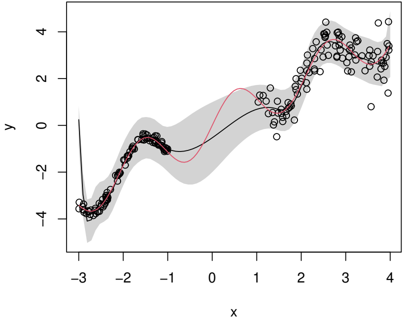
$$\mu(x) = \frac{\sum_{j=1}^J a_j(x)\, h_j\, \mu_j(x)}{\sum_{j=1}^J a_j(x)\, h_j}$$

- Parameters $(w_j, \gamma_j, \beta_j, \alpha_j, \sigma_j^2, h_j)$ learned from training data

ENNreg, cont.

- Left: from Sec 4.1 of Denoeux's *BELIEF'22* paper
- Right: conformal prediction IM on a "similar" data set





Concluding remarks

- Today: regression & imprecise-probabilistic methods
- Conformal prediction is powerful, but has limitations:
  - computationally expensive
  - exchangeability isn't always an appropriate assumption[10]
  - marginally but not conditionally valid
  - efficiency gains by replacing "$\sup \mathsf{P}$" with a general "$\overline{\mathsf{P}}$"?
- Random fuzzy numbers are new and promising
  - seems quite flexible
  - is it provably valid...?

[10] e.g., Mao, M., and Reich, `arXiv:2006.15640`

Next lecture

- Formal decision theory
- Precise-probabilistic version
  - von Neumann–Morganstern and others
  - maximize expected utility
- Imprecise-probabilistic version
  - Choquet integrals define lower/upper expected utility
  - how to optimize?
- Applications