ST790 — Fall 2022
Imprecise-Probabilistic Foundations of Statistics

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

Week 11a

This lecture

- Multinomial model
- Prediction of a categorical response
- A few different approaches[1]
  - gBayes based on Walley's *Imprecise Dirichlet Model*
  - Denoeux's confidence-region-based belief function
  - my new prediction IM
- ...

[1] Not an exhaustive list, e.g., conformal prediction can be applied here too

Multinomial model

- Consider a set of $K \geq 2$ categories
  - could be ordered (e.g., small, medium, large)
  - could be unordered (e.g., red, blue, green)
- Let X denote a random variable on $\mathbb{X} = \{1, 2, \ldots, K\}$
- Distribution P of X determined by a probability vector

$$\theta_k = \mathrm{P}(X = k), \quad k = 1, \ldots, K$$

- All three are equivalent:
  - parameter space $\mathbb{T}$ for $\theta = (\theta_1, \ldots, \theta_K)$
  - set of all probability distributions P for X
  - probability simplex in $\mathbb{R}^K$

Multinomial model, cont.

- Let $X^n = (X_1, \ldots, X_n)$ be iid copies of X
- Likelihood function is

$$L_n(\theta) \propto \prod_{k=1}^{K} \theta_k^{N_k}, \quad N_k(X^n) = |\{i : X_i = k\}|$$

- This likelihood is "nonparametric" by above equivalence
- For inference on θ (equivalently, on P):
  - maximum likelihood, $\hat{\theta}_k = N_k / n$
  - Bayes, e.g., with Dirichlet prior (below)
  - ...
- Our goal here is predicting a new observation, X_{n+1}

Dirichlet distribution

- $\mathrm{Dir}_K(\beta)$: continuous distribution on the simplex $\mathbb{T} \subset \mathbb{R}^K$
- Density function,[2] depending on $\beta = (\beta_1, \ldots, \beta_K)$,

$$\vartheta \mapsto c(\beta) \prod_{k=1}^{K} \vartheta_k^{\beta_k - 1}, \quad \vartheta \in \mathbb{T}$$

- It's the Bayesian conjugate prior for multinomial models[3]
  - if $\Theta \sim \mathrm{Dir}_K(\beta)$
  - and $(X^n \mid \Theta = \theta) \overset{\text{iid}}{\sim} \mathrm{Mult}_K(\theta)$
  - then $(\Theta \mid X^n = x^n) \sim \mathrm{Dir}_K(\beta + N(x^n))$
  - and the predictive distribution is

$$\mathrm{P}(X_{n+1} = k \mid x^n) = \frac{\beta_k + N_k(x^n)}{\sum_{\kappa=1}^{K} \beta_\kappa + N_\kappa(x^n)}, \quad k = 1, \ldots, K$$

[2] https://en.wikipedia.org/wiki/Dirichlet_distribution

[3] This is the basis for Ferguson's Dirichlet process developments

Walley's imprecise Dirichlet model

- The Bayesian analysis above depends on the choice of β
- If information about β is available, then fine
- If not, then what? A "default" choice?
- Walley[4] aimed to be more careful by allowing the Dirichlet prior to be *imprecise*, i.e., a set of Dirichlet priors
- Reparametrization of the Dirichlet model:
  - mean vector $t = (t_1, \ldots, t_K) \in \mathbb{T}$ and precision $s > 0$
  - then $\beta_k = st_k$, for $k = 1, \ldots, K$
- Walley proposed a prior credal set

$$\mathscr{C}(s) = \{\mathrm{Dir}_K(s, t) : t \in \mathbb{T}\}$$

- Almost vacuous...

[4] Walley (*JRSS-B* 1996), "Learning about a bag of marbles"

IDM, cont.

- By the conjugacy result above, the posterior credal set is

$$\mathscr{C}(x^n; s) = \{\mathrm{Dir}_K(s+n, t^n) : t_k^n = \tfrac{st_k + N_k(x^n)}{s+n},\ t \in \mathbb{T}\}$$

- This is a set of posterior distributions for Θ, so gBayes inference on Θ calculates lower/upper envelopes
- Our goal is prediction of X_{n+1}, and the above credal set for a collection of predictive distributions indexed by (s, x^n)
- Read off the lower/upper prediction probabilities:

$$\underline{\Pi}_{x^n,s}(A) = \frac{\sum_{k \in A} N_k(x^n)}{s+n}$$

$$\overline{\Pi}_{x^n,s}(A) = \frac{s + \sum_{k \in A} N_k(x^n)}{s+n}, \quad A \subseteq \mathbb{X}$$

IDM, cont.

- Imprecision is controlled by the precision s
  - large s means wider spacing between $\underline{\Pi}_{x^n,s}$ and $\overline{\Pi}_{x^n,s}$
  - small s means narrower spacing
  - can interpret s as a "learning rate"
- Properties:
  - super-simple to implement
  - it's generalized Bayes, so entirely coherent
  - output is a belief function in this case[5]
  - both $\underline{\Pi}_{x^n,s}$ and $\overline{\Pi}_{x^n,s}$ converge to true P as $n \to \infty$
  - ...
- Walley considered the multinomial model specifically, but these ideas extend to other exponential families[6]

[5] $m(\{k\}) = N_k/(n+s)$, for $k = 1, \ldots, K$, and $m(\mathbb{X}) = s/(n+s)$

[6] e.g., Quaeghebeur & de Cooman (*ISIPTA* 2005)

Denoeux's belief function

- Thierry Denoeux is a leader in the belief function community, fundamental work on stat inference & ML
- A really nice paper[7] of his is on the construction of a belief function for predicting $X_{n+1} \sim \text{Mult}_K(\cdot)$
- Background:
  - I showed you Dempster's framework for $K = 2$ (binomial)
  - a belief function for prediction follows readily
  - computationally challenging for $K \geq 3$...[8]
  - very recent work[9] helps to overcome this challenge
- Denoeux's paper gives a relatively simple alternative to Dempster's approach for general K

[7] Denoeux (*IJAR* 2006)
[8] Dempster (*Ann Math Stat* 1966)
[9] Jacob et al (*JASA* 2021)

Denoeux's belief function, cont.

- Goal: a belief function $\underline{\Pi}_{X^n}$ on $\mathbb{X}$ for predicting/quantifying uncertainty about the next observation X_{n+1}
- Lots of options, need some properties we want $\underline{\Pi}_{X^n}$ to satisfy
- Denoeux's two requirements:
  - R1 $\underline{\Pi}_{X^n}(A) \to \mathsf{P}(A)$ in P-probability, all $A \subseteq \mathbb{X}$, as $n \to \infty$
  - R2 For a given $\alpha \in (0, 1)$,

$$\mathsf{P}\{\underline{\Pi}_{X^n}(A) \leq \mathsf{P}(A) \text{ for all } A\} \geq 1 - \alpha$$

- Both are reasonable
- Superficially at least, R2 looks similar to *validity*, but it's actually very different; more later
- How to find $\underline{\Pi}_{X^n}$ that satisfies R1 and R2?

Denoeux's belief function, cont.

- Recall:
  - the multinomial parameter $\theta = (\theta_1, \ldots, \theta_K)$
  - equivalance between θ and P
- A $100(1-\alpha)\%$ confidence region $C_\alpha(X^n)$ for θ satisfies

$$\mathsf{P}\{C_\alpha(X^n) \ni \theta\} \geq 1-\alpha$$

- For $C_\alpha(X^n)$, Denoeux recommends
  - $C_\alpha(X^n) = [\theta_1^-, \theta_1^+] \times \cdots \times [\theta_K^-, \theta_K^+]$
  - where

$$\theta_k^\pm = \frac{a + 2N_k \pm \Delta_k^{1/2}}{2(n+a)}$$

  - with $a = \mathtt{qchisq}(1-\alpha, \mathtt{df} = 1)$ and

$$\Delta_k = a\Big\{a + \frac{4N_k(n-N_k)}{n}\Big\}$$

Denoeux's belief function, cont.

- Each θ in $C_\alpha(X^n)$ corresponds to a probability dist on $\mathbb{X}$
- Lower envelope defines a candidate solution

$$\underline{\Pi}_{X^n}^{\text{tmp}}(A) = \max\Big\{ \sum_{k \in A} \theta_k^-, 1 - \sum_{k \notin A} \theta_k^+ \Big\}, \quad A \subseteq \mathbb{X}$$

- Properties:
  - easy to check R1, $n^{-1} N_k(X^n) \to \theta_k = \mathrm{P}(X = k)$
  - similarly for R2, i.e.,

$$\mathrm{P}\{\underline{\Pi}_{X^n}^{\text{tmp}} \text{ lower-bounds } \mathrm{P}\} = \mathrm{P}\{C_\alpha(X^n) \ni \theta(\mathrm{P})\} \geq 1 - \alpha$$

  - note that $\underline{\Pi}_{X^n}$ depends on α...
- *However*, $\underline{\Pi}_{X^n}^{\text{tmp}}$ isn't a belief function[10] when $K > 3$

[10] But it is a 2-monotone capacity...

Denoeux's belief function, cont.

- Denoeux wants the output to be a belief function, so he needs to modify $\underline{\Pi}_{X^n}^{\text{tmp}}$ in a suitable way
- Natural idea: inner approximation of $\underline{\Pi}_{X^n}^{\text{tmp}}$ by a belief function
- This approximation is more complicated, requires optimization via solving a linear program
- Too messy to present here, but apparently easy to do
- Denoeux shows that the output, $\underline{\Pi}_{X^n}$, of this optimization routine is a belief function and satisfies R1 and R2

Valid prediction IM

- In discrete settings, nonparametric = parametric
- Can do the IM stuff from before with multinomial model
- If I take a vacuous prior, then

$$\pi_{x^n}(\kappa) = \sup_{\theta} \mathsf{P}_{X^n, X_{n+1} \mid \theta}\{\eta(X^n, X_{n+1}) \le \eta(x^n, \kappa)\}$$

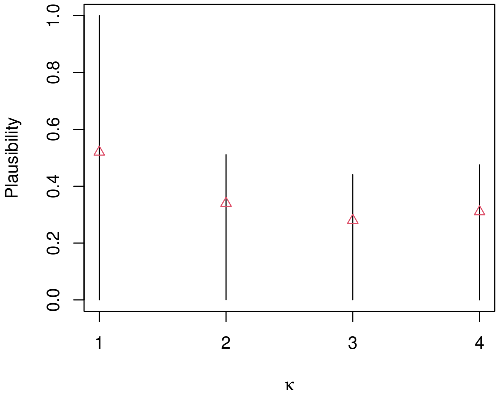
  where

$$\eta(x^n, \kappa) = \frac{\sup_\theta \theta_\kappa^{N_\kappa + 1} \prod_{k \neq \kappa} \theta_k^{N_k}}{\max_\zeta \sup_\theta \theta_\zeta^{N_\zeta + 1} \prod_{k \neq \zeta} \theta_k^{N_k}}$$

- Looks messier than it really is...

Illustration

- Data from Denoeux's Example 1: $N(x^n) = (91, 49, 37, 43)$
- Plot shows my mine and Denoeux's ($\triangle$) plausibility contour



Discussion

- Multinomial models are simple but represent an important class of problems — these are "discrete nonparametric"
- Walley's IDM is simple and powerful[11]
- Denoeux's method is appealing:
  - very simple in the $K \in \{2, 3\}$ cases
  - doable but more complicated in others
  - motivated by some performance-related criteria
  - ...
- Denoeux's R2 is not the same as "validity"
- I threw the IM solution together quickly and naively, would be interesting to explore this further...

[11] Extensive literature on this covering pros and cons

Next lecture

- Prediction with covariates
  - regression
  - classification
- Imprecise probabilistic methods
- ...