Supervised
Posterior Distributions

Peter Grünwald
CWI Amsterdam, www.grunwald.nu

Petri Kontkanen, Petri Myllymäki, Teemu Roos
Henry Tirri, Hannes Wettig
Complex Systems Computation Group (CoSCo),
Helsinki Institute for Information Technology

Overview

1. A Small Problem
2. A Small Solution
3. A **BIG** Problem
4. Small Solution solves Big Problem!

Overview

1. A Small Problem
   - Parametric inference for conditional prediction
2. A Small Solution
   - 'Supervised' posterior distributions
3. A **BIG** Problem
   - Model Selection/Averaging for cond. prediction
4. Small Solution solves Big Problem!

Setup

- Data: $D = ((x_1, y_1), \ldots, (x_n, y_n))$

  where $x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \{0, 1\}$

- Model: $\mathcal{M} = \{P_{X,Y}(\cdot|\theta) \mid \theta \in \Gamma\}$

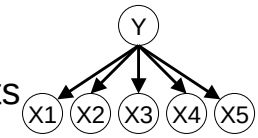
  – parametric i.i.d. model of joint distribution

Task: Conditional Prediction

- Infer ('estimate') distribution θ from D
- Use θ to predict future values of Y given future values of X
- Measure quality of prediction using some loss function
- Examples:
  - **classification** loss
  - conditional log score

Task: Conditional Prediction

- **Example:**
  - $\mathcal{M}$ is Naïve Bayes model:
  - $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_m$
  - $\mathcal{X}$ consists of m components with finite ranges



$$p(x_1, \ldots, x_m, y|\theta) = p(y|\theta_y) \prod_{i=1}^{m} p(x_i|y, \theta_{x_i|y})$$

Empirical Observation

- Using maximum *conditional* ('partial') likelihood estimator

$$\hat{\theta}_{Y|X} := \arg\max_{\theta \in \Gamma} \log p(y^n | x^n, \theta) = \arg\max_{\theta \in \Gamma} \sum_{i=1}^{n} \log p(y_i | x_i, \theta)$$

'often' leads to (much) 'better' inferences than the ordinary ('full') ML estimator

$$\hat{\theta}_{X,Y} := \arg\max_{\theta \in \Gamma} \log p(x^n, y^n | \theta)$$

Empirical Observation

- 'better' means both
  - smaller conditional logarithmic score

$$\sum_{j=n+1}^{k} \log p(y_j | x_j, \hat{\theta}_{Y|X}) \gg \sum_{j=n+1}^{k} \log p(y_j | x_j, \hat{\theta}_{X,Y})$$

  - smaller classification (0/1) loss

…on a test set of data from the same source

Discriminative vs Generative

- Discriminative models: CML = ML
  - e.g. regression, Bayes net with only incoming arcs at Y
- Generative ('sampling') models: CML better
  - especially for large sample sizes
  - also outperforms Bayesian posterior
  - holds for most types of models for which it was tried
    - *Nearly always* the case for Naïve Bayes if sample large enough
    - Many recent papers e.g. by Greiner & Zhou, Kontkanen et al., Jebara, Friedman, Geiger, Goldszmidt, many others
    - Less recent papers: e.g. Dawid '76

Small Problem: Bayes!

- The Bayesian in me is waking up:

If inference based on conditional ML is good, then inference based on conditional posterior is even better!

Small Problem: Bayes!

- The Bayesian in me is waking up:

If inference based on conditional ML is good, then inference based on ~~conditional~~ posterior is even better!

'SUPERVISED'

Small Problem: Bayes!

- Problem: how to define 'supervised' posterior?
- Idea: analogously to

$$p(\theta \mid x^n, y^n) \propto p(x^n, y^n \mid \theta) p(\theta)$$

we want

$$p_{\text{super}}(\theta \mid x^n, y^n) \propto p(y^n \mid x^n, \theta) p(\theta)$$

Solution – Simple Version

- Define the supervised version $\mathcal{M}_{\mathrm{super}}$ of $\mathcal{M}$ as follows:

$$p_{\mathrm{super}}(y|x,\theta) := p(y|x,\theta)$$

$$p_{\mathrm{super}}(y^n|x^n,\theta) := \prod_{i=1}^{n} p(y_i|x_i,\theta)$$

- $p_{\mathrm{super}}(x|\theta)$ remains undefined, hence $\mathcal{M}_{\mathrm{super}}$ is a conditional model.

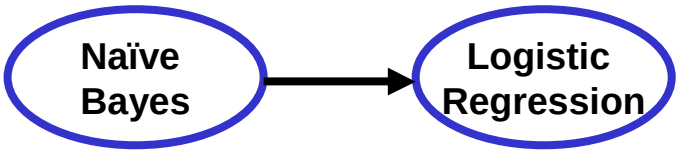
Solution

- To indicate why this solves our problem, note that by Bayes' rule, for any fixed x^n we have for all y^n :

$$\begin{aligned} p_{\mathrm{super}}(\theta|x^n,y^n) &= \frac{p_{\mathrm{super}}(y^n|x^n,\theta)p(\theta)}{\sum_{y^n} p_{\mathrm{super}}(y^n|x^n,\theta)p(\theta)} \\ &\propto p(y^n|x^n,\theta)p(\theta) \end{aligned}$$

Solution

- We effectively turned a sampling (generative) model into a diagnostic (discriminative) model
- for example:



- in this form identical to earlier proposals, e.g. Heckerman & Meek's (1997) BERC-models

Overview

1. A Small Problem
   - Parameter estimation for conditional prediction
2. A Small Solution
   - 'Supervised' posterior distributions
3. **A BIG Problem**
   - Model Selection/Averaging for cond. pred.
4. Small Solution solves Big Problem!

The BIG Problem: Model Selection/Averaging

- Example: $\mathcal{M}_1, \ldots, \mathcal{M}_k$
  - Bayesian network models over
    $$X_1 \times \ldots X_m \times Y$$
    with different structure (DAG)
  - to be used for classification/conditional prediction

The BIG Problem

- With uniform prior over models,
  $$p(\mathcal{M}_1) = \ldots = p(\mathcal{M}_k) = \frac{1}{k}$$

…. Bayesian model selection picks model with maximum 'score'…

$$\mathbf{S}_{\mathrm{marglik}}(x^n, y^n | \mathcal{M}) = \log p(x^n, y^n | \mathcal{M})$$

…which is just the log-marginal likelihood

The BIG Problem

- In analogy to previous findings, large marginal ('full, unconditional') likelihood may not be a good indicator of the quality of conditional predictions based on the model
- Empirical observations/experiments with artificially generated data confirm this in striking manner!
  - Robert Cowell (AI & Stats, 2001) can tell you all about it!

The BIG Problem

- In analogy to conditional Maximum Likelihood, may want to base Bayesian model selection on conditional Marginal Likelihood:

$$\mathbf{S}_{\text{cond-marglik}}(x^n, y^n | \mathcal{M}) = \log p(y^n | x^n, \mathcal{M})$$

- Suggested by several researchers
  - e.g. Buntine '93, Jebara (website),…

2 Problems:

1. Experiments suggest that conditional marginal likelihood

$$\mathbf{S}_{\text{cond-marglik}}(x^n, y^n | \mathcal{M})$$

does not work very well for model selection for classification! (Kontkanen et al. '99)

2 Problems:

2. Ordinary Bayesian model selection has prequential interpretation (Dawid '84,'91):

$$\log p(x^n, y^n | \mathcal{M}) = \sum_{i=1}^{n} \log p(x_i, y_i | (x^{i-1}, y^{i-1}), \mathcal{M})$$

but conditional marginal likelihood does not:

$$\log p(y^n | x^n, \mathcal{M}) \neq \sum_{i=1}^{n} \log p(y_i | x_i, (x^{i-1}, y^{i-1}), \mathcal{M})$$

$\mathbf{S}_{\text{cond-marglik}}(x^n, y^n | \mathcal{M})$ $\quad$ $\mathbf{S}_{\text{cond-preq}}(x^n, y^n | \mathcal{M})$

The Solution

CLAIM: The 'supervised posterior' leads to a natural definition of model score

$$\mathbf{S}_{\text{super}}(x^n, y^n|\mathcal{M}) = \log p_{\text{super}}(y^n|x^n, \mathcal{M}) =$$
$$\log \int \prod p(y_i|x_i, \theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta$$

which is in some sense 'better' than both the conditional marginal likelihood and the conditional prequential method

Asymptotic Analysis

- Suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d. $\sim P^*$, where P^* not necessarily in $\mathcal{M}$
- Asymptotic behaviour of diverse scores

$\mathbf{S}_{\text{cond-marglik}}(x^n, y^n|\mathcal{M}), \mathbf{S}_{\text{cond-preq}}(x^n, y^n|\mathcal{M}), \mathbf{S}_{\text{super}}(x^n, y^n|\mathcal{M})$

under regularity conditions on $\mathcal{M}$

– satisfied for Bayesian network structures

Theorem 1, well-specified case

- If $P^* \in \mathcal{M}$ then with P^*- probability 1,

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{S}_{\text{cond-marglik}}(X^n, Y^n|\mathcal{M}) =$$
$$\lim_{n\to\infty} \frac{1}{n}\mathbf{S}_{\text{cond-preq}}(X^n, Y^n|\mathcal{M}) =$$
$$\lim_{n\to\infty} \frac{1}{n}\mathbf{S}_{\text{super}}(X^n, Y^n|\mathcal{M}) =$$
$$E_{P^*}[\log P^*(Y|X)]$$

Theorem 2, misspecified case

- Let:

$$\tilde{\theta}^{(j)}_{Y|X} := \arg\min_{\theta\in\Gamma_{\mathcal{M}_j}} E_{P^*}[-\log p(Y|X,\theta,\mathcal{M}_j)]$$

$$\tilde{\theta}^{(j)}_{X,Y} := \arg\min_{\theta\in\Gamma_{\mathcal{M}_j}} E_{P^*}[-\log p(X,Y|\theta,\mathcal{M}_j)]$$

Theorem 2, misspecified case

- If $P^* \notin \mathcal{M}$ then asymptotically,
  - $\mathbf{S}_{\text{super-I}}$ selects $\mathcal{M}_j$ achieving

$$\min_j E_{P^*}[-\log p(Y|X, \tilde{\theta}^{(j)}_{Y|X}, \mathcal{M}_j)]$$

WHICH IS OPTIMAL!

  - $\mathbf{S}_{\text{cond-preq}}$ selects $\mathcal{M}_j$ achieving

$$\min_j E_{P^*}[-\log p(Y|X, \tilde{\theta}^{(j)}_{X,Y}, \mathcal{M}_j)]$$

REASONABLE BUT IN GENERAL SUBOPTIMAL

  - $\mathbf{S}_{\text{cond-marglik}}$ does something

QUITE UNREASONABLE!

Conclusion & Future Work

- **SUMMARY:** 'natural' definition of supervised posterior for parameter estimation, prediction; showed its potential in model selection/ averaging in conditional prediction/ classification settings
- **FUTURE WORK:**

Try this on real-world data sets!

Thank you for your attention!