Robust Probabilistic Modeling with Bayesian Data Reweighting

https://github.com/yixinwang/robust-rpm-public

本論文では、確率モデルをミスラベルに対して頑健にするためのレシピを示す。

方法：
ガウス尤度をより裾野の重いt分布で置き換える、など。
(これより洗練された確率モデルを我々は使うけどね〜)

提案手法概要

reweighted probabilistic models (RPM)(再重み付け確率モデル?)

1. まず確率モデルを作る
2. それぞれの観測データの貢献度を調整する。（それぞれの観測データの尤度を使う？）
3. 元の確率モデルの潜在変数と共にこれらの重みを推論した。

この調整された確率モデルの事後分布は観測がモデルの仮説と整合するかを判断してくれる。

定式化

N個の独立な観測$y = (y_1, y_2, \ldots .y_N)$を考える。尤度関数は観測データごとの尤度の積にできる。

$$p(y|\beta) = \prod_{n=1}^{N} l(y_n|\beta)$$

(yの尤度＝yを観測する確率)

ここでβはyの分布を決める潜在変数。事前確率を$p_\beta(\beta)$とする。

Bayesian data reweightingの３ステップについて

1. 確率モデル$p(y, \beta)$を定義

$$p(y, \beta) = p_\beta(\beta) \prod_{n=1}^{N} l(y_n|\beta)$$

2. 正の潜在変数w_nを使って各観測の尤度を増幅する。同時にその$w = (w_1, w_2, \ldots, w_N)$の事前分布$p_w(w)$を選択する。

$$
\begin{aligned}
p(y, \beta, w) &= p(w)p(y, \beta | w) \\
&= \frac{1}{Z} p_\beta(\beta) p_w(w) \prod_{n=1}^{N} l(y_n | \beta)^{w_n}
\end{aligned}
$$

   これがreweighted probabilistic model (RPM)

3. 潜在変数βとwの事後分布$p(\beta, w | y)$を推定する。

重みwにより、RPMはどの観測がその仮定と一致し(つまり本当に正しいラベルである)、どの観測が一致しない(つまりミスラベル)かを自動的に探索することができる。

ざっくりいうと、事後分布推定はRPMの対数をとった、

$$
\log p_\beta(\beta) + \log p_w(w) + \sum_{n=1}^{N} w_n \log l(y_n | \beta) \tag{1}
$$

こいつをβとwについて最大化することである。
ここでwの事前分布$p_w(w)$がcorrupted dataを当てるのに重要な役目を果たすので、Section2で三つほど試してみた。

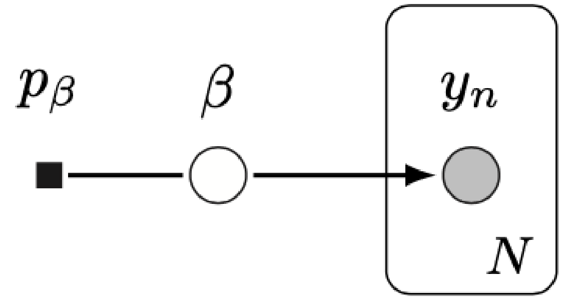
まず、重みwが式(1)に与える影響について。 w_nを小さくすると、$w_n \log l(y_n | \beta)$の項は増えるが、$\log w_n$は小さくなる。
y_nがcorrupted dataの場合は$\log l(y_n | \beta)$が非常に小さくなるので、w_nはより大きい値として推定されます。

潜在変数βが式 (1) に与える影響について。 corruptedな観測データの重みw_nが小さくなるにつれて、尤度項は、corrupted dataに低質量を割り当て、データセットの残りの部分に焦点を当てることができる。重みと潜在変数が連携して、起こりそうにない測定値を自動的に識別し、元のモデルの仮定に一致する観測に焦点を合わせる。

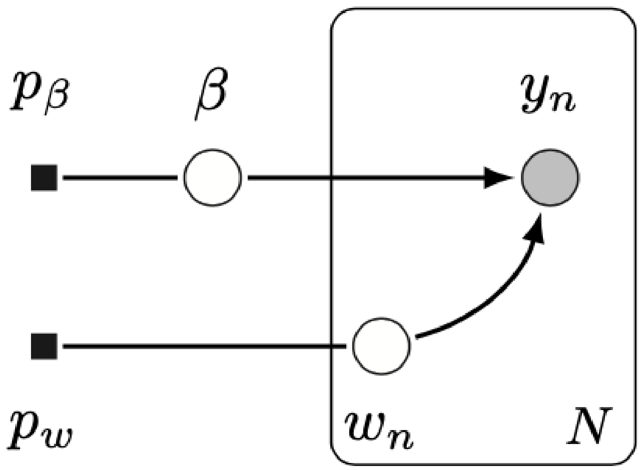
Related Work

1. 特定の確率モデルを、不正確な観測に対してより頑健にすることに焦点を当てている研究
2. ロバストベイジアン分析？
   （事前分布の作り方に対する感度分析を行った？）
3. データのreweighting(元々はアンサンブルのためだったけど、各データの重要度信頼度の推定にも使えそう？)

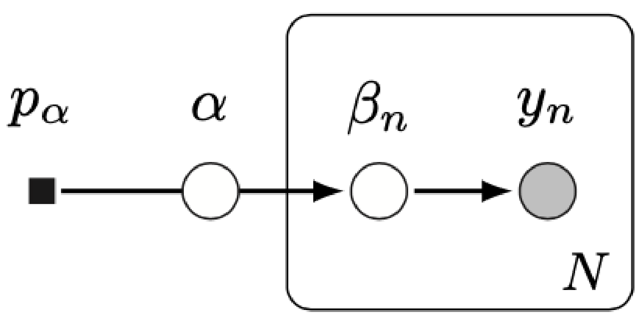
グラフィカルモデル



(a) Original probabilistic model



(b) Reweighted probabilistic model (RPM)



(c) Localized probabilistic model

Figure 2. RPMs begin with a probabilistic model **(a)** and introduce a set of weights w as latent variables. This gives a model **(b)** that explores which data observations match its assumptions. Localization **(c)**, instead, builds a hierarchical model. (Appendix A shows when a localized model is also an RPM.)

RPM

尤度$l(\cdot|\beta)$がLesbegue基底測度を持つ指数分布族mに属する時、正規化係数Zは1に収束する？

ようわからんけど、wの事前分布である$p_w(w)$が結果に影響を及ぼすから大事と言っている。なので、ベータ分布、スケーリングされたディリクレ分布、ガンマ分布で事前分布を張って調べてみた。

ベータ分布

ここでは$w_n \in (0,1)$としている。それぞれの重みは事前分布では独立同分布に従うとする。

$$p_w(w) = \prod_{n=1}^{N} Beta(w_n; a, b) \tag{4}$$

これはRPMにとって最も保守的な選択肢。wを考えない元のモデルよりもピークが大きくなることはない。

スケーリングされたディリクレ分布

ここではw_nの合計がNとなるようにしている。

$$v = \frac{w}{N}$$
$$p_v(v) = Dirichlet(a\mathbf{1}) \tag{5}$$

$\mathbf{1}$は$N \times 1$の1が並んだベクトル。wを考えない元のモデルでは全てのwが1であると考えることができる。ハイパラのaを小さく設定すれば重みwの変更がされやすくなるし、大きくするとディリクレ分布は滑らかになる。(あんま変更されない)

ガンマ分布

各重みw_nは独立同分布。

$$p_w(w) = \prod_{n=1}^{N} Gamma(w_n; a, b) \tag{6}$$

wの値が無限に大きくなりうるのであんましおすすめしない。

実験からはベータ分布が良さそう、という話になった。

理論的なお話

事前分布をガンマ分布としたときのw_nのMAP推定量を計算してみると、

$$\hat{w}_n = \frac{a-1}{b - \log l(y_n|\beta)} \tag{8}$$

となる。$a > 1$の時$\hat{w}_n$はlに対して単調増加。だからlが小さい、つまりラベルが怪しいと$\hat{w}_n$も小さくなり、そのあとのβの推定に寄与するデータが選別されていく。

後半はようわからんかった...

事後分布の推定について

$p(\beta, w|y)$を求めたい。でも簡単に式で描けるような形にはならない。なので、確率的プログラミングシステムStanで自動推論を用いた。

あと次の観測$y_\dagger$を得る確率分布は$p(y_\dagger|y)$で表されるが、これが次の$l(\cdot)$となるし、判別問題を解く学習済みのモデルとなるわけである。RPMなら、

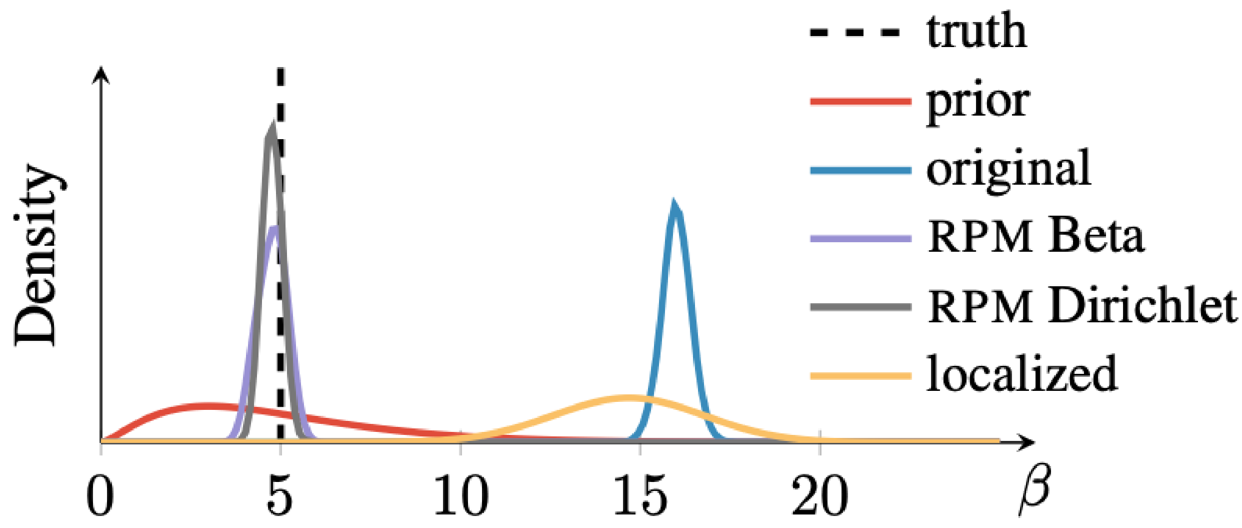
$$p_{RPM}(y_\dagger|y) = \int\int p(y_\dagger|\beta, w_\dagger)p(\beta|y)p(w_\dagger)dw_\dagger d\beta$$

実験

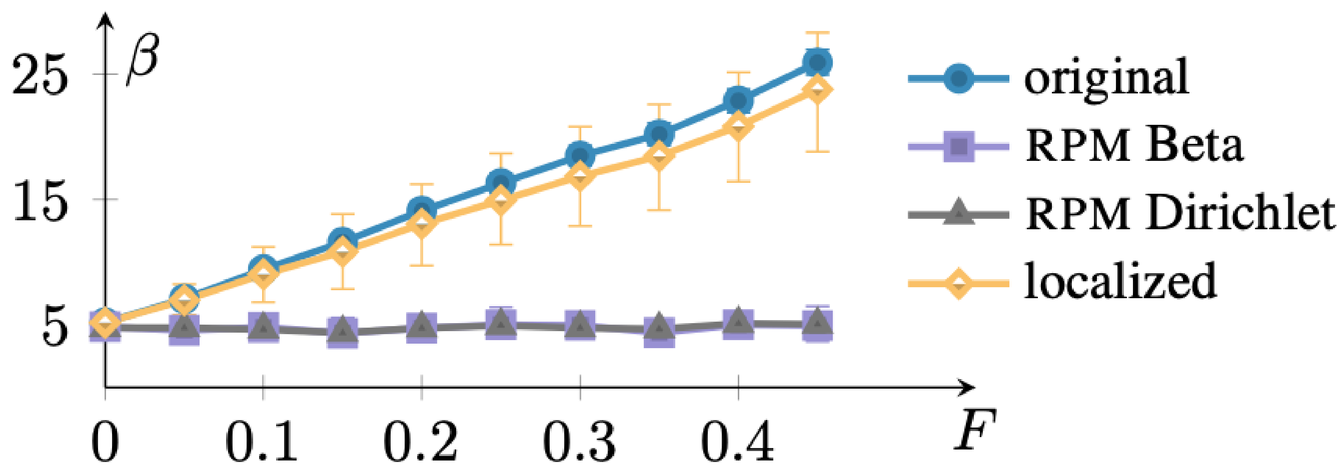
人口データと実データ(レコメンドのデータ)でやったよ。

ルータがパケットを受け取るまでの時間(ミスラベル)

が観測できている時、この時間は$Po(\beta)$に従い、$\beta \sim G(a, b)$とする。人口データとして、正しい分布は$\beta = 5$とし、でも$F\%$のパケットが失敗して$\beta = 50$のポアソン分布にしたがってルータまで届くとしよう。w_nの事前分布$G(a = 2, b = 0.5)$でやった結果が図3。



(a) Posteriors for $F = 25\%$ failure rate.



(b) Posterior 95% credible intervals.

Figure 3. Outliers simulation study. We compare $\text{Beta}(0.1, 0.01)$ and $\text{Dir}(\mathbf{1})$ as priors for the reweighted probabilistic model. **(a)** Posterior distributions on β show a marked difference in detecting the correct wait-time rate of $\beta = 5$. **(b)** Posterior 95% confidence intervals across failure rates F show consistent behavior for both Beta and Dirichlet priors. ($N = 100$ with 50 replications.)

ちゃんとRPMでモデル化した奴はFが大きくなっても安定してるね！

で、この結果からベータ分布とディリクレ分布そんなに変わらなかったから、我々はベータ分布を採用したんだze。

色盲者の推定(潜在変数のミス)

色覚異常の状態と家族歴に関する男女のデータセットがある。性別の情報はないよ。
通常色盲は男性がなりやすいんだけど、そのことを我々は知らないとしよう。

色盲かそうでないかを表す二値ラベルy_nは当然ベルヌーイ分布に従う。

$$y_n \sim Bernoulli(\frac{1}{1+\exp(-p_n)})$$

人口データを作る時、
p_nは、$p/n \sim \beta X$でサンプルされる。ただし男性なら$\beta = 0.5$、女性なら$\beta = 0.01$からサンプルされるようにしておく。ここでXが家族歴を表す数値で、$X \sim U(-10, 10)$。

この時、男性としてサンプルされたデータy_nが正常ラベル、女性としてサンプルされたデータy_nはミスラベルとする。

$$p_n = \beta x_n$$

これでデータセット$\{x_n, y_n\}_{n=1}^{N}$が与えられた時に、しっかりクラス分類できるでしょうか...!?
(ここでβの事前分布を$\mathcal{N}(0, 10)$、w_nの事前分布を$Beta(0.1, 0.01)$とする。)

同じく$F\%$のサンプルを女性にして人口データを作ってやった結果がこちら。

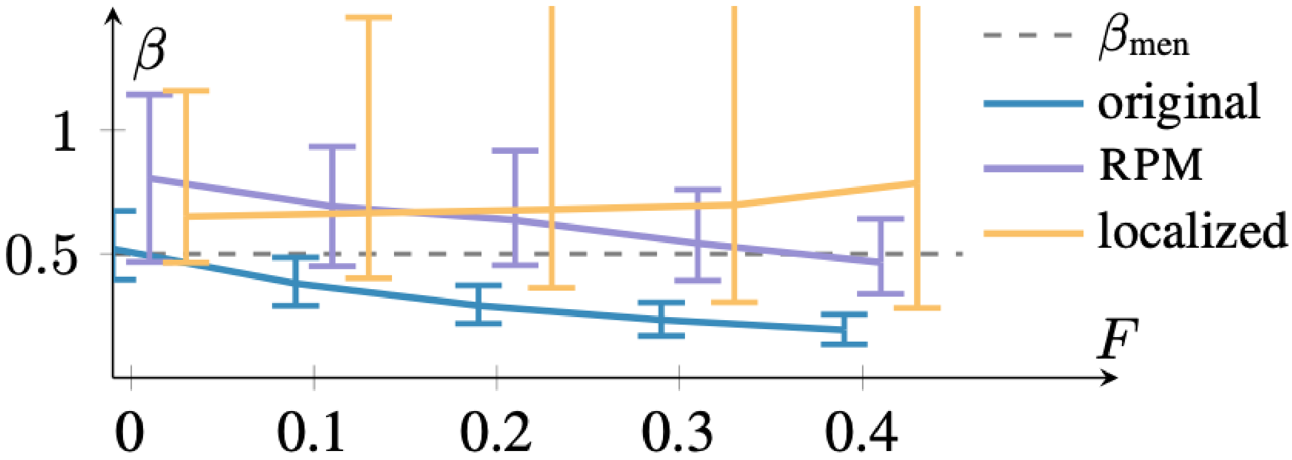


Figure 5. Missing latent groups study. Posterior 95% credible intervals for the RPM always include the dominant $\beta_{men} = 0.5$, as we vary the percentage of females in the data. Dataset size $N = 100$ with 50 replications.

あ〜ハイハイちゃんとβが0.5の周りにいますね。(でもRPMなんかFが大きいほど0.5に近づいてない....?)

追記

> RPMは、現実との不一致を検出する診断ツールとしても機能する。推定した重みの分布は、元のモデルの仮定を無視している観測データの存在を示す。図6は推定後の重みのカーネル密度推定値を示す。ミスラベルのない人口データセットは、1に近い重みを受ける。対照的に、ミスラベルのある実際のデータセットは、重みの二峰性分布を示した。

肺癌リスクのお話(モデル化のミス)

肺癌になるかどうかを当てたい。タバコを吸っているなら確実に癌になる可能性は高いが、肥満であるならどうだろう？そもそもタバコと肥満も相互に関係し合っているらしい。

タバコの使用量をx_1、肥満度をx_2としよう。人口データを今回は三種類の方法で作った。それがTable 1。
例えば一行目なら、

$$y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

として作ったデータセット$\{x_{1n}, x_{2n}, y_n\}_{n=1}^{N}$から推定したい、しかしモデル化する時に

$$y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

としちゃってもちゃんとβ_1を当てられるんかな？ってのをみたい。(実際は50回推定した時の絶対偏差を小さくしたい)