Making News Go Viral

Math 640 Class Project Presentation

Cin Cin Fang & Michael Leibert
(updated: 2019-05-09)

Data

Online News Popularity Data Set

- This dataset contains rougly 45 attributes about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).
- Variables range from category (business, entertainment, lifestyle, etc), measures of the article's subjectivity and polarity, and the average word length.
- We end up removing some variables that display multicollinearity, and we also scale the continuous variables.

Data

A sample of a few rows and columns:

```
##                                                       url num_imgs
##  http://mashable.com/2013/07/03/lion-forge-80s-tv-comics/        6
##           http://mashable.com/2013/07/03/low-cost-iphone/       15
##          http://mashable.com/2013/07/03/mediashift-wi-fi/        1
##
##  num_videos global_subjectivity max_negative_polarity topic shares
##           0         0.421512501                -0.025  tech    792
##           1         0.503344852                 -0.05 other 843300
##           0         0.301536797                  -0.1 world   9500
##
```

- Dataset has 39,644 samples. We will use a n=30,000 train / 9,644 test split.
- There are 25 covariates, with $K = 30$ estimated parameters, including the intercept.

Goals

- The goal is to find the attributes that can predict how sharable an article is.
- We will utilize the Bayesian ordinal probit model.
- We will test the robustness of our estimates for β with three different priors on β: a flat prior, a Laplacian prior, and a Normal prior.
- The Laplacian and Normal priors turn our regression into Bayesian versions of LASSO regression and ridge regression respectively, a purposeful choice as we have many covariates under consideration in our regression.
- We will use out-of-sample testing to check our results.

Methods

- Rather than attempt to predict the given "shares" directly, we instead translate "shares" of the articles $i = 1, \ldots, n$ into quintiles to form ordinal categories.

- Again, this lends itself to ordinal probit regression, which can be more easily modeled in a Bayesian context through latent variables.

- For this formulation, the latent variable z_i of article i, $z \overset{iid}{\sim} N(x_i'\beta, 1)$
  - Where x_i is the vector of predictive variables and β is the vector of coeffcients we are ultimately interested in.

The observed response variable y_i, which quantile of shares does it attract, is:

$$y_i = \begin{cases} 1 & \text{if } \gamma_0 < z_i \leq \gamma_1 \\ 2 & \text{if } \gamma_1 < z_i \leq \gamma_2 \\ & \vdots \\ 5 & \text{if } \gamma_4 < z_i \leq \gamma_5 \end{cases}$$

Where γ_0 is $-\infty$, γ_1 is fixed at 0, and γ_5 is ∞

Flat Prior

Likelihood

$$\mathcal{L}(y|z,\gamma,\beta,X) \propto \prod_{i=1}^{n}\left[\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\, I_{\{y_i=j\}}\right]\exp\left[-\frac{1}{2}\left(z_i - x_i'\beta\right)^2\right]$$

$$\propto \left[\prod_{i=1}^{n}\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\, I_{\{y_i=j\}}\right]\exp\left[-\frac{1}{2}(z - X\beta)'\,(z - X\beta)\right]$$

Under the flat prior $\pi(\beta) \propto 1$, the posterior is unchanged.

$$P(z,\gamma,\beta|y,X) \propto \left[\prod_{i=1}^{n}\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\, I_{\{y_i=j\}}\right]\exp\left[-\frac{1}{2}(z - X\beta)'\,(z - X\beta)\right]$$

Flat Prior

We now derive the conditional for β:

$$p(\beta|z,\gamma,X,y) \propto \exp\left[-\frac{1}{2}\left(\beta'X'X\beta - 2\beta'X'X(X'X)^{-1}X'z\right)\right]$$

Which is the kernel of a normal distribution, so

$$\beta|z,\gamma,X,y \sim N\Big((X'X)^{-1}X'z,(X'X)^{-1}\Big)$$

The conditional for γ_j, where $j = 2,3,4$:

$$p(\gamma_j|z,\beta,X,y) \propto \prod_{i=1}^{n} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\, I_{\{y_i=j\}} + I_{\{\gamma_j<z_i\leq\gamma_{j+1}\}}\, I_{\{y_i=j+1\}}$$

Which is a uniform distribution

$$\gamma_j|z,\beta,X,y \sim U\Big(\max\{\max\{z_i : y_i = j\},\gamma_{j-1}\},\min\{\min\{z_i : y_i = j+1\},\gamma_{j+1}\}\Big)$$

Flat Prior

To find the conditionals for z_i, we must consider the separate cases of y_i

If $y_i = j$, the conditional for z_i is:

$$p(z|\gamma, \beta, X, y) \propto \exp\left[-\frac{1}{2}\left(z_i - x_i'\beta\right)^2\right] I_{\{\gamma_{j-1} < z_i \leq \gamma_j\}}$$

Which is a truncated normal with mean $x_i'\beta$, variance 1, and truncated to be between γ_{j-1} and γ_j.

Flat Prior

We specifiy a Gibbs sampling scheme:

1. For each $j = 2, \ldots, J$ draw $\gamma_j^{(b)}$ from

$$U\left(\max\left\{\max\left\{z_i^{(b-1)} : y_i = j\right\}, \gamma_{j-1}^{(b-1)}\right\}, \min\left\{\min\left\{z_i^{(b-1)} : y_i = j+1\right\}, \gamma_{j+1}^{(b-1)}\right\}\right)$$

2. For each $j = 1, \ldots, J$ draw $z_i^{(b)} | y_i = j$ from

$$N\left(x_i'\beta^{(b-1)}, 1\right), \text{truncated at the left (right) by } \gamma_{j-1}^{(b-1)} \left(\gamma_j^{(b-1)}\right)$$

3. Draw $\beta^{(b)}$ from

$$N\left(\left(X'X\right)^{-1} X'z^{(b)}, \left(X'X\right)^{-1}\right)$$

Penalization

- With a large number of predictors our estimates can be noisy.
- We wish to perform regularization to give more stable estimates and perform shrinkage.
- We will consider the Bayesian Lasso and a Ridge Regression in an attempt to produce better estimates.
- Recall the Bayesian Lasso uses the double exponential prior and the Bayesian Ridge regression utilizes a normal prior.

L1 LASSO

Likelihood

$$\begin{aligned}\mathcal{L}(y|z,\gamma,\beta,X) &\propto \prod_{i=1}^{n}\left[\ \sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\ I_{\{y_i=j\}}\right]\exp\left[-\frac{1}{2}\big(z_i - x_i'\beta\big)^2\right] \\ &\propto \left[\ \prod_{i=1}^{n}\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\ I_{\{y_i=j\}}\right]\exp\left[-\frac{1}{2}(z-X\beta)'\,(z-X\beta)\right]\end{aligned}$$

L1 LASSO

Posterior

- Instead of applying the Laplacian prior to directly to the likelihood, we can express the Laplacian as a mixture model of normals with inverse gamma priors.

- Previously: $\beta_k \sim L(0, \lambda^{-1})$

- Now: $\beta_k \sim N\left(0, \frac{4\left(\lambda^{-1}\right)^2}{\alpha_k}\right)$, where $\left(\lambda^{-1}\right)^2 \sim IG(a, b)$ and $\alpha_k \overset{iid}{\sim} IG\left(1, \frac{1}{2}\right)$

- This allows us to utilize a Gibbs sampler

L1 LASSO

Full Posterior

$$P(z,\gamma,\beta,\alpha,\lambda|X,y) \propto \mathcal{L}(y|z,\gamma,\beta,X)\pi(\beta|\lambda,\alpha)\pi(\lambda)\pi(\alpha)$$

$$\propto \left[\prod_{i=1}^{n}\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\, I_{\{y_i=j\}}\right] \exp\left[-\frac{1}{2}(z-X\beta)'\,(z-X\beta)\right]\left[\left(\lambda^{-1}\right)^2\right]^{-(a+1)}.$$

$$\exp\left(-\frac{b}{\left(\lambda^{-1}\right)^2}\right)\prod_{k=1}^{K}\left(\frac{\alpha_k}{\left(\lambda^{-1}\right)^2}\right)^{\frac{1}{2}}\exp\left[-\frac{\alpha_k\,\beta_k^2}{8\left(\lambda^{-1}\right)^2}\right]\alpha_k^{-2}\exp\left(-\frac{1}{2\alpha_k}\right)$$

L1 LASSO

The full conditionals of z and γ_j are not affected by the new prior.

Recall :

If $y_i = j$, the conditional for z_i is:

$$p(z|\gamma,\beta,\alpha,\lambda,X,y) \propto \exp\left[-\frac{1}{2}\left(z_i - x_i'\beta\right)^2\right] I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}$$

A truncated normal with mean $x_i'\beta$, variance 1, and truncated to be between γ_{j-1} and γ_j.

The conditional for γ_j is:

$$\gamma_j|z,\beta,X,y \sim U\Big(\max\{\max\{z_i : y_i = j\},\gamma_{j-1}\},\min\{\min\{z_i : y_i = j+1\},\gamma_{j+1}\}\Big)$$

L1 LASSO

The conditional for β is:

$$P(\beta|\text{rest}) \propto \exp\left\{-\frac{1}{2}\left[\left(\beta-\left(X'X+\frac{\lambda^2}{4}\mathbf{D}_\alpha\right)X'Z\right)'\left(X'X+\frac{\lambda^2}{4}\mathbf{D}_\alpha\right)\left(\beta-\left(X'X+\frac{\lambda^2}{4}\mathbf{D}_\alpha\right)X'Z\right)\right]\right\}$$

Where $\mathbf{D}_\alpha = \mathbf{diag}(\alpha_1, \ldots, \alpha_k)$

We recognize this as the kernel of a normal distribution, so

$$\beta|\text{rest} \sim N\left(\left(X'X+\frac{\lambda^2}{4}\mathbf{D}_\alpha\right)^{-1}X'z, \left(X'X+\frac{\lambda^2}{4}\mathbf{D}_\alpha\right)^{-1}\right)$$

L1 LASSO

The conditional on λ is:

$$P(\lambda|z,\gamma,\beta,\alpha,X,y) \propto \left(\lambda^2\right)^{\frac{K}{2}+a+1} \exp\left[-\lambda^2\left(b+\frac{1}{8}\sum_{k=1}^{K}\alpha_k\beta_k^2\right)\right]$$

Which is identifiable as the kernel of a Gamma.

$$\lambda|z,\gamma,\beta,\alpha,X,y \sim Gamma\left(\frac{K}{2}+a+2, b+\frac{1}{8}\sum_{k=1}^{K}\alpha_k\beta_k^2\right)$$

L1 LASSO

Finally, the conditional for α_p

$$P(\alpha_p|\lambda, z, \gamma, \beta, \alpha_{p\neq k}, X, y) \propto \alpha_p^{-\frac{3}{2}} \exp\left[-\frac{1}{2}\frac{\left(\alpha_p - \frac{2\lambda^{-1}}{|\beta_p|}\right)^2}{\alpha_p\left(\frac{2\lambda^{-1}}{|\beta_p|}\right)^2}\right]$$

Which is a kernel of the inverse Gaussian.

$$\alpha_p|\lambda, z, \gamma, \beta, \alpha_{p\neq k}, X, y \sim N^{-1}\left(\frac{2\lambda^{-1}}{|\beta_k|}, 1\right)$$

L1 LASSO

After deriving the conditionals we can set up a Gibbs sampler without a M-H step.

1. For each $j = 2, \ldots, J$ draw $\gamma_j^{(b)}$ from

$$U\left(\max\left\{\max\left\{z_i^{(b-1)} : y_i = j\right\}, \gamma_{j-1}^{(b-1)}\right\}, \min\left\{\min\left\{z_i^{(b-1)} : y_i = j+1\right\}, \gamma_{j+1}^{(b-1)}\right\}\right)$$

2. For each $j = 1, \ldots, J$ draw $z_i^{(b)} | y_i = j$ from

$$N\left(x_i'\beta^{(b-1)}, 1\right), \text{truncated at the left (right) by } \gamma_{j-1}^{(b-1)} \left(\gamma_j^{(b-1)}\right)$$

3. Draw $\lambda^{2^{(b)}}$ from $Gamma\left(\frac{K}{2} + a + 2, b + \frac{1}{8}\sum_{k=1}^{K} \alpha_k^{(b-1)} \beta_k^{2^{(b-1)}}\right)$

4. For each $k = 1, \ldots, K$ draw $\alpha_k^{(b)}$ from $N^{-1}\left(\frac{2\lambda^{-1^{(b)}}}{|\beta_k^{(b-1)}|}, 1\right)$

5. Draw $\beta^{(b)}$ from

$$N\left(\left(X'X + \frac{\lambda^{2^{(b)}}}{4}\mathbf{D}_\alpha^{\mathbf{(b)}}\right)^{-1} X'z^{(b)}, \left(X'X + \frac{\lambda^{2^{(b)}}}{4}\mathbf{D}_\alpha^{\mathbf{(b)}}\right)^{-1}\right)$$

L2 Ridge regression

Lastly, we consider the Bayesian Ridge regression, which is $\beta_k \sim N\left(0, \lambda^{-1}\right)$. Thus we have:

$$
\begin{aligned}
\pi\left(\beta_1, \ldots, \beta_K\right) &\propto \prod_{k=1}^{K} \lambda^{\frac{1}{2}}\ \exp\left[-\frac{\lambda}{2}\beta_k^2\right] \\
&\propto \lambda^{\frac{K}{2}} \exp\left[-\frac{\lambda}{2}\sum_{k=1}^{K}\beta_k^2\right]
\end{aligned}
$$

With Jeffrey's prior as the hyper-prior:

$$\pi(\lambda) \propto \lambda^{-1}$$

L2 Ridge regression

Same likelihood as before.

This leads to the posterior:

$$
\begin{aligned}
P(\beta,\lambda,z|y) &\propto \mathcal{L}(y|\beta,\lambda,z)\pi(\beta|\lambda)\pi(\lambda) \\
&\propto \left[\ \prod_{i=1}^{n}\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\ I_{\{y_i=j\}}\right]\exp\left[-\frac{1}{2}(z-X\beta)'\,(z-X\beta)\right]\cdot \\
&\qquad \exp\left[-\frac{1}{2}(z-X\beta)'(z-X\beta)\right]\lambda^{\frac{K}{2}}\exp\left[-\frac{\lambda}{2}\sum_{k=1}^{K}\beta_k^2\right]\lambda^{-1} \\
&\propto \left[\ \prod_{i=1}^{n}\sum_{j=1}^{5} I_{\{\gamma_{j-1}<z_i\leq\gamma_j\}}\ I_{\{y_i=j\}}\right]\lambda^{\frac{K}{2}-1}\exp\left[-\frac{1}{2}\Big((z-X\beta)'(z-X\beta)+\lambda\beta'\beta\Big)\right]
\end{aligned}
$$

L2 Ridge regression

Thus the full conditional on λ is:

$$P(\lambda|\text{rest}) \propto \lambda^{\frac{K}{2}-1} \exp\left[-\frac{\lambda}{2}\ \beta'\beta\right]$$

Which is recognizable as a $Gamma\left(\frac{K}{2}, \frac{1}{2}\beta'\beta\right)$

The conditional of β is:

$$P(\beta|\text{rest}) \propto \exp\left[-\frac{1}{2}\left(\beta'(X'X+\lambda I_K)\beta - 2\beta(X'X+\lambda I_K)(X'X+\lambda I_K)^{-1}X'z\right)\right]$$

Which is the kernel of a multivariate normal random variable with distribution

$$\beta|\text{rest} \sim N\left((X'X+\lambda I_K)^{-1}X'z, (X'X+\lambda I_K)^{-1}\right)$$

L2 Ridge regression

The full conditionals of z and γ_j are not affected by the new prior.

Recall :

If $y_i = j$, the conditional for z_i is:

$$p(z|\text{rest}) \propto \exp\left[-\frac{1}{2}\left(z_i - x_i'\beta\right)^2\right] I_{\{\gamma_{j-1} < z_i \leq \gamma_j\}}$$

A truncated normal with mean $x_i'\beta$, variance 1, and truncated to be between γ_{j-1} and γ_j.

The conditional for γ_j is:

$$\gamma_j|\text{rest} \sim U\Big(\max\{\max\{z_i : y_i = j\}, \gamma_{j-1}\}, \min\{\min\{z_i : y_i = j+1, \gamma_{j+1}\}\}\Big)$$

L2 Ridge regression

We finally specifiy a Gibbs sampling scheme

1. For each $j = 2, \ldots, J$ draw $\gamma_j^{(b)}$ from

$$U\left(\max\left\{\max\left\{z_i^{(b-1)} : y_i = j\right\}, \gamma_{j-1}^{(b-1)}\right\}, \min\left\{\min\left\{z_i^{(b-1)} : y_i = j+1\right\}, \gamma_{j+1}^{(b-1)}\right\}\right)$$

2. For each $j = 1, \ldots, J$ draw $z_i^{(b)} | y_i = j$ from

$$N\left(x_i'\beta^{(b-1)}, 1\right), \text{truncated at the left (right) by } \gamma_{j-1}^{(b-1)} \left(\gamma_j^{(b-1)}\right)$$

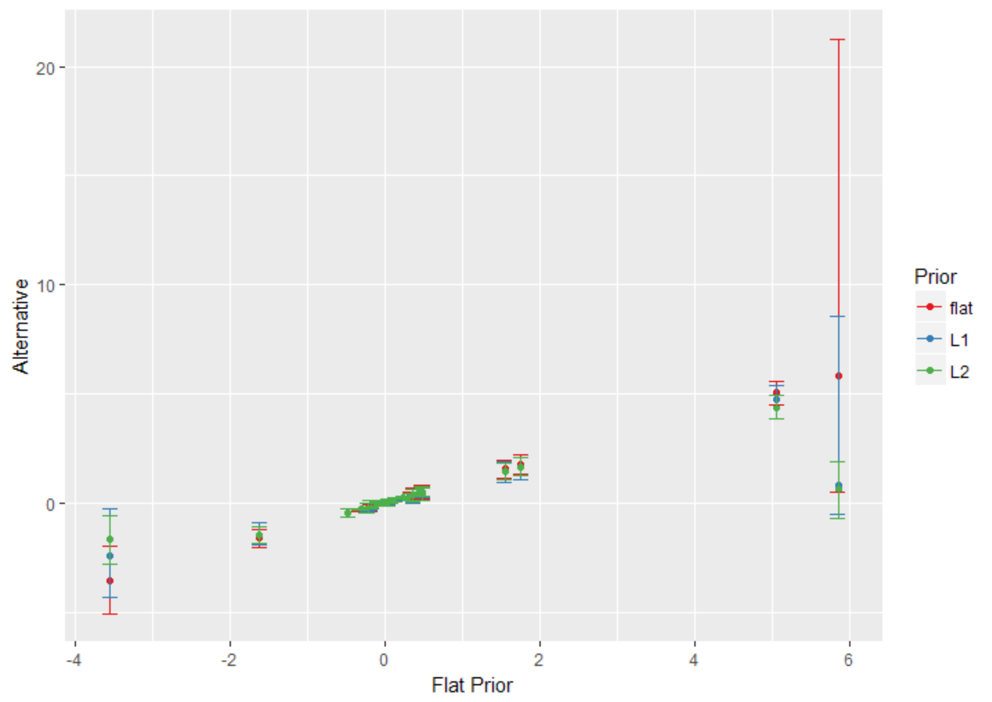
3. Draw $\lambda^{(b)}$ from $Gamma\left(\frac{K}{2}, \frac{1}{2}\beta'^{(b-1)}\beta^{(b-1)}\right)$

4. Draw $\beta^{(b)}$ from

$$N\left(\left(X'X + \lambda^{(b)} I_k\right)^{-1} X'z^{(b)}, \left(X'X + \lambda^{(b)} I_k\right)^{-1}\right)$$

Results

- We sample using the data in a training set consisting of 30,000 datapoints.
- The parameters of interest are the coeffcients of the regression, β, and the thresholds for the categories, γ.
- Our estimates of some of the parameters β clearly are affected by the prior:



Results

The variables that are associated with increased sharability are `kw_avg_avg` (average shares of the average keyword), `num_hrefs` (number of links), `self_reference_avg_sharess` (average shares of referenced articles in Mashable), while `global_rate_negative_words` and `num_self_hrefs` (number of links to other articles published by Mashable) is associated with less shares.

This suggests the topic is more important than the style, except that Mashable audiences are not looking for negative articles.

The flat prior also does worse than random guessing in an out-of-sample test, while the Bayesian LASSO and ridge models do marginally better.

Prior	Accuracy
Flat Prior	**0.180**
Laplacian Prior	**0.224**
Normal Prior	**0.222**

In all three models, the predictions are mostly 5's, showing that our models tend to over-predict sharability.

Results

Confusion Matrix for Flat Prior

Truth	Estimated 2	Estimated 3	Estimated 4	Estimated 5
1	10	14	23	1893
2	11	17	22	2144
3	15	11	17	1836
4	4	16	29	1835
5	4	16	42	1685

Confusion Matrix for L1 Prior

1	116	931	288	605
2	94	839	376	885
3	51	504	395	929
4	32	379	406	1067
5	29	263	297	1158

Confusion Matrix for L2 Prior

1	31	928	806	175
2	18	822	1024	330
3	22	475	969	413
4	7	364	992	521
5	8	271	811	657