mml87_nb

Kelvin

27 June 2017

MLE of NB parameters

The MLE estimations of NB parameters are -*as follows:

$$\begin{aligned} P(Y=y) &= \frac{count(y)}{n} \\ P(X_j=x|Y=y) &= \frac{count(x)\cup count(y)}{count(y)} \end{aligned}$$

Negative log likelihood of NB

For a Naive Bayes model with binary variable, its parameters are $\{P(Y_i=T), P(X_{ij}|Y_i=T), P(X_{ij}|Y_i=F)\}$. To simply the notations, we use $\{p_{i0}, p_{ij1}, p_{ij2}\}$ to denote the above probabilities respectively. The likelihood of Naive Bayes given a data set is then

$$l = \prod_i^n P(Y_i=T|\vec{X}_i)^{Y_i}(1-P(Y_i=T|\vec{X}_i))^{1-Y_i}$$

where the posterior probability of Y_i given a vector $\vec{X}_i = < X_{i1}, \dots, X_{im} >$ is

$$\begin{aligned} P(Y_i=T|\vec{X}_i) &= \frac{P(Y_i=T)\prod_{j=1}^m P(X_{ij}|Y_i=T)}{P(\vec{X}_i)} \\ &= \frac{P(Y_i=T)\prod_{j=1}^m P(X_{ij}|Y_i=T)}{P(Y_i=T)\prod_{j=1}^m P(X_{ij}|Y_i=T) + (1-P(Y_i=T))\prod_{j=1}^m P(X_{ij}|Y_i=F)} \\ &= \frac{p_{i0}\prod_{j=1}^m p_{ij1}}{p_{i0}\prod_{j=1}^m p_{ij1} + (1-p_{i0})\prod_{j=1}^m p_{ij2}} \end{aligned}$$

The negative loglikelihood

$$\begin{aligned} L &= -\sum_{i=1}^n \left[Y_i \ln p(Y_i|\vec{X}_i) + (1-Y_i)\ln(1-p(Y_i|\vec{X}_i)) \right] \\ &= -\sum_{i=1}^n \left[Y_i \ln p_{i0} + (1-Y_i)\ln(1-p_{i0}) + Y_i \sum_{j=1}^m \ln p_{ij1} + (1-Y_i)\sum_{j=1}^m \ln p_{ij2} - \ln\left(p_{i0}\prod_{j=1}^m p_{ij1} + (1-p_{i0})\prod_{j=1}^m p_{ij2} \right) \right] \end{aligned}$$

Fisher information matrix

The first derivatives of the above negative log likelihood w.r.t. each parameter are

$$
\begin{aligned}
\frac{\partial L}{\partial p_{i0}} &= -\sum_{i=1}^{n}\left[\frac{Y_i}{p_{i0}} - \frac{1-Y_i}{1-p_{i0}} - \frac{\prod_{j=1}^{m} p_{ij1} - \prod_{j=1}^{m} p_{ij2}}{p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}}\right] \\
\frac{\partial L}{\partial p_{ik1}} &= -\sum_{i=1}^{n}\left[\frac{Y_i}{p_{ik1}} - \frac{p_{i0}\prod_{j=1}^{m} p_{ik1}}{p_{ik1}\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)}\right] \\
\frac{\partial L}{\partial p_{ik2}} &= -\sum_{i=1}^{n}\left[\frac{1-Y_i}{p_{ik2}} - \frac{(1-p_{i0})\prod_{j=1}^{m} p_{ik2}}{p_{ik2}\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)}\right]
\end{aligned}
$$

The second derivatives are

$$
\begin{aligned}
\frac{\partial^2 L}{\partial p_{i0}^2} &= \sum_{i=1}^{n}\left[\frac{Y_i}{p_{i0}^2} + \frac{1-Y_i}{(1-p_{i0})^2} - \left(\frac{\prod_{j=1}^{m} p_{ij1} - \prod_{j=1}^{m} p_{ij2}}{p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}}\right)^2\right] \\
\frac{\partial^2 L}{\partial p_{ik1}} &= \sum_{i=1}^{n}\left[\frac{Y_i}{p_{ik1}^2} - \left(\frac{p_{i0}\prod_{j=1}^{m} p_{ij1}}{p_{ik1}\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)}\right)^2\right] \\
\frac{\partial^2 L}{\partial p_{ik2}} &= \sum_{i=1}^{n}\left[\frac{1-Y_i}{p_{ik2}^2} - \left(\frac{(1-p_{i0})\prod_{j=1}^{m} p_{ij2}}{p_{ik2}\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)}\right)^2\right] \\
\frac{\partial^2 L}{\partial p_{i0}p_{ik1}} &= \sum_{i=1}^{n}\frac{\prod_{j=1}^{m} p_{ij1}p_{ij2}}{\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)^2}\frac{1}{p_{ik1}} \\
\frac{\partial^2 L}{\partial p_{i0}p_{ik2}} &= \sum_{i=1}^{n}\frac{\prod_{j=1}^{m} p_{ij1}p_{ij2}}{\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)^2}\frac{-1}{p_{ik2}} \\
\frac{\partial^2 L}{\partial p_{ik1}p_{ik2}} &= \sum_{i=1}^{n}\frac{\prod_{j=1}^{m} p_{ij1}p_{ij2}}{\left(p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}\right)^2}\frac{-p_{i0}(1-p_{i0})}{p_{ik1}p_{ik2}}
\end{aligned}
$$

Since FIM entries are expectations of the second derivatives, we need to take expectations for the first three second derivatives that contain Y_i. To simplify the notations, we use p_x to denote $p_{i0}\prod_{j=1}^{m} p_{ij1} + (1-p_{i0})\prod_{j=1}^{m} p_{ij2}$. Then the expectations become

$$
\begin{aligned}
E\left(\frac{\partial^2 L}{\partial p_{i0}^2}\right) &= \sum_{i=1}^{n}\left[\frac{\prod_{j=1}^{m} p_{ij1}}{p_{i0}p_x} + \frac{\prod_{j=1}^{m} p_{ij2}}{(1-p_{i0})p_x} - \left(\frac{\prod_{j=1}^{m} p_{ij1} - \prod_{j=1}^{m} p_{ij2}}{p_x}\right)^2\right] \\
E\left(\frac{\partial^2 L}{\partial p_{ik1}}\right) &= \sum_{i=1}^{n}\left[\frac{p_{i0}\prod_{j=1}^{m} p_{ij1}}{p_{ik1}^2 p_x} - \left(\frac{p_{i0}\prod_{j=1}^{m} p_{ij1}}{p_{ik1}p_x}\right)^2\right] \\
E\left(\frac{\partial^2 L}{\partial p_{ik2}}\right) &= \sum_{i=1}^{n}\left[\frac{(1-p_{i0})\prod_{j=1}^{m} p_{ij2}}{p_{ik2}^2 p_x} - \left(\frac{(1-p_{i0})\prod_{j=1}^{m} p_{ij2}}{p_{ik2}p_x}\right)^2\right]
\end{aligned}
$$

MML of Naive Bayes

$$I = -\ln K - \ln h(\vec{\theta}) + \frac{1}{2}\ln F(\vec{\theta}) - \ln f(D|\vec{\theta}) + \frac{|\vec{\theta}|}{2}$$

where $\vec{\theta} =< p_{i0}, p_{ij1}, p_{ij2} >, \forall j \in [1, m]$ is the set of parameters, $|\vec{\theta}|$ is the number of free parameters, K is the lattice contant and $h(\vec{\theta})$ is the parameter prior. A commonly used conjugate prior for binary variables is beta prior (i.e., beta distribution) with probability density function

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. For simplicity, we assume all parameters are uniformly (i.e., $\alpha = \beta = 1$), hence for a single parameter prior is $x(1-x)$. Assuming all parameters are independent, we have

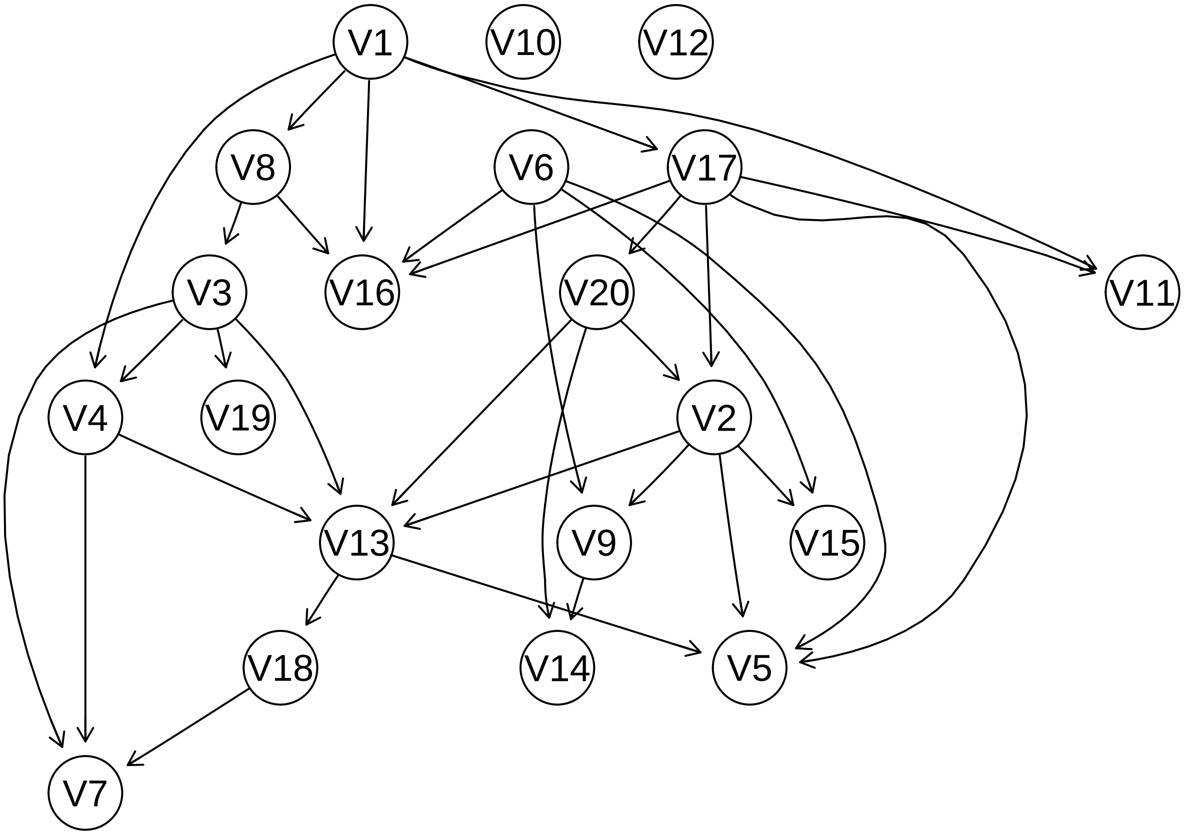
$$\ln h(\vec{\theta}) = \ln p_{i0} + \ln(1-p_{i0}) + \sum_{j=1}^{m} [\ln p_{ij1} + \ln(1-p_{ij1}) + \ln p_{ij2} + \ln(1-p_{ij2})]$$

Substituting this into the above MML formula we get

$$I = -\left(\ln p_{i0} + \ln(1-p_{i0}) + \sum_{j=1}^{m} [\ln p_{ij1} + \ln(1-p_{ij1}) + \ln p_{ij2} + \ln(1-p_{ij2})]\right) + \frac{1}{2}F(\vec{\theta}) - \ln f(D|\vec{\theta}) + \frac{d}{2}(1+\ln k_d)$$

where d is the number of free parameters and k_d is the lattice constant for each free parameter.

Random model



Executing MML_NB

MML_CPT

```
## Search: Forward greedy with mmlCPT
## 0 parent: 673.4999
## parents = V1 : 669.1356
## parents = V8 : 675.8705
## parents = V17 : 409.5802
## parents = V20 : 650.1703
## parents = V2 : 664.1456
## parents = V3 : 675.7342
## parents = V4 : 676.0923
## parents = V6 : 673.8693
## parents = V13 : 676.203
## parents = V18 : 676.0859
## parents = V5 : 669.7929
## parents = V7 : 675.8306
## parents = V16 : 657.119
## parents = V9 : 664.9188
## parents = V14 : 674.987
## parents = V19 : 675.067
## parents = V10 : 674.2535
## parents = V15 : 675.1523
## parents = V12 : 676.1391
## add V17 into mb
## current mb is { V17 } with msg len 409.5802
## -------------------------------
## parents = V17 V1 : 393.2037
## parents = V17 V8 : 411.1062
## parents = V17 V20 : 414.7114
## parents = V17 V2 : 413.1693
## parents = V17 V3 : 413.9034
## parents = V17 V4 : 414.3343
## parents = V17 V6 : 410.4308
## parents = V17 V13 : 414.3358
## parents = V17 V18 : 414.2848
## parents = V17 V5 : 414.8275
## parents = V17 V7 : 413.9837
## parents = V17 V16 : 412.9111
## parents = V17 V9 : 412.9861
## parents = V17 V14 : 411.91
## parents = V17 V19 : 414.3912
## parents = V17 V10 : 411.7188
## parents = V17 V15 : 412.5888
## parents = V17 V12 : 414.9861
## add V1 into mb
## current mb is { V17 V1 } with msg len 393.2037
## -------------------------------
## parents = V17 V1 V8 : 399.3168
## parents = V17 V1 V20 : 398.6399
## parents = V17 V1 V2 : 395.2637
## parents = V17 V1 V3 : 398.8307
## parents = V17 V1 V4 : 399.9947
## parents = V17 V1 V6 : 394.4156
## parents = V17 V1 V13 : 399.0746
## parents = V17 V1 V18 : 399.7728
## parents = V17 V1 V5 : 400.0038
## parents = V17 V1 V7 : 399.983
## parents = V17 V1 V16 : 399.3498
## parents = V17 V1 V9 : 394.996
## parents = V17 V1 V14 : 395.9365
## parents = V17 V1 V19 : 399.2849
## parents = V17 V1 V10 : 395.3892
## parents = V17 V1 V15 : 397.0728
## parents = V17 V1 V12 : 398.9956
## Stop! No better choice for MB!
```

```
## [1] "V17" "V1"
```

```
## [1] 703.6197
```

The above mml + cpt is even smaller than mml + nb, which is expected to be shorter due to less number of parameters comparing with a full cpt!!!

MML_Logit

```
## $mml
## [1] 771.8173
##
## $nlogPrior
## [1] 11.14519
##
## $nlogLattice
## [1] -6.544965
##
## $logF
## [1] 26.9223
##
## $nll
## [1] 753.756
```

```
## Search: Forward greedy with mmlLogit
## 0 parent: 673.7474
## parents = V1 : 669.157
## parents = V8 : 675.9686
## parents = V17 : 409.1885
## parents = V20 : 650.2334
## parents = V2 : 664.23
## parents = V3 : 675.8381
## parents = V4 : 676.1918
## parents = V6 : 684.4484
## parents = V13 : 676.3008
## parents = V18 : 676.1822
## parents = V5 : 669.8772
## parents = V7 : 675.9283
## parents = V16 : 657.0965
## parents = V9 : 664.9732
## parents = V14 : 675.082
## parents = V19 : 675.1715
## parents = V10 : 674.137
## parents = V15 : 675.2622
## parents = V12 : 676.2402
## add V17 into mb
## current mb is { V17 } with msg len 409.1885
## ------------------------------
## parents = V17 V1 : 1369.017
## parents = V17 V8 : 947.9792
## parents = V17 V20 : 411.1952
## parents = V17 V2 : 411.1174
## parents = V17 V3 : 410.1375
## parents = V17 V4 : 411.3593
## parents = V17 V6 : 415.2702
## parents = V17 V13 : 411.037
## parents = V17 V18 : 410.5204
## parents = V17 V5 : 411.3026
## parents = V17 V7 : 411.0448
## parents = V17 V16 : 411.0758
## parents = V17 V9 : 410.9265
## parents = V17 V14 : 408.9776
## parents = V17 V19 : 410.6045
## parents = V17 V10 : 409.4339
## parents = V17 V15 : 410.0302
## parents = V17 V12 : 411.3511
## add V14 into mb
## current mb is { V17 V14 } with msg len 408.9776
## ------------------------------
## parents = V17 V14 V1 : 1449.452
## parents = V17 V14 V8 : 888.8437
## parents = V17 V14 V20 : 414.3453
## parents = V17 V14 V2 : 413.3804
## parents = V17 V14 V3 : 425.5867
## parents = V17 V14 V4 : 415.4749
## parents = V17 V14 V6 : 2256.36
## parents = V17 V14 V13 : 420.737
## parents = V17 V14 V18 : 422.4788
## parents = V17 V14 V5 : 418.3312
## parents = V17 V14 V7 : 412.8379
## parents = V17 V14 V16 : 419.0797
## parents = V17 V14 V9 : 420.3545
## parents = V17 V14 V19 : 410.3802
## parents = V17 V14 V10 : 409.2096
## parents = V17 V14 V15 : 409.9942
## parents = V17 V14 V12 : 411.1271
## Stop! No better choice for MB!
```

```
## [1] "V17" "V14"
```

Sanity check

This is a sanity check to ensure that nll is corrected calculated. The following code use gRain to compute the posterior probability $P(Y_i|\vec{X}_i)$ based on the estimated cpts from data. The posterior probability given each data point is then used to compute the nll of the entire data set. The answer confirms that the above nll calculation is correct.

Comparison

Observations

- The above tests show that nll for both naive bayes and logit are close to nll using true cpts. But mml_nb is larger than mml_logit and mml_cpt even if the true model is a naive bayes model. This is likely to be caused by large message length for nb parameter priors and definitely large log(det(fim)) as compared with mml_logit. Could this because of the mml_logit is only for 1st order logit model, hence the model is simpler than a nb?
- In mml_nb, det(fim) are almost always negative when X is a single variable. When Xs are two or more variables, it is less likely to have negative determinant. Not sure what's the problem. This problem may have been fixed. But occasionally the determinant is still a small negative number, perhaps due to under flow.