Sparse Prior

August 20, 2013

1 Sparse Prior

We introduce sparse prior to improve the prediction accuracy of our Naive Bayes classifier. We assign each $\theta_{y,i}^{(k)}(j)$ with a binary random variable $Z_{y,i}^{(k)}(j)$. $\theta_{y,i}^{(k)}(j) = 0$ when $Z_{y,i}^{(k)}(j) = 0$. Fix y, k, j, $\{\theta_{y,i}^{(k)}(j) : Z_{y,i}^{(k)}(j) = 1\}$ are sampled from Dirichlet distribution as before.

1.1 the Model

Fix y, k, denote i as position index for feature vector, and $j = \{1, 2, \ldots, 8\}$ as class index, thus

$$\begin{aligned}
\theta_{i,j} &:= \theta_{y,i}^{(k)}(j) \\
Z_{i,j} &:= Z_{y,i}^{(k)}(j)
\end{aligned}$$

$$\begin{aligned}
Z_{i,j} &\sim Bernoulli(p_i) \\
\theta_{i,j} &= \frac{\mu_{i,j} Z_{i,j}}{\sum_{j'=1}^{8} \mu_{ij'} Z_{ij'}} \\
\mu_{i,j} &\sim Gamma(\alpha_{i,j}, 1)
\end{aligned}$$

In order to simulate posterior distribution of θ_{ij}, we need to simulate joint distribution $\mathbb{P}(p_i, Z_{ij}, \mu_{ij} | Data)$. We use Gibbs sampler to simulate this distribution.

1.2 Gibbs Sampler

Fix y, k. Let X be a data matrix, and each row represents a feature vector x and $Y(x) = y$, the conditional distributions for this model are

1. $\mu_{ij}, \forall i, j | Z_{i'j'}, \forall i'j', p_{i'}, \forall i', X$
   Fix i,

   - For j with $Z_{ij} = 0$, simulate μ_{ij} from prior distribution.

- For j with $Z_{ij} = 1$, let $\mathcal{J} = \{j : Z_{ij} = 1\}$. First we can simulate $\{\frac{\mu_{ij}}{\sum_{j' \in \mathcal{J}} \mu_{ij'}} : j \in \mathcal{J}\}$ given data, and this is a Dirichlet distribution with parameters ($\alpha_{ij} +$ # of occurrence of class j for feature i : $j \in \mathcal{J}$). Given $\{\frac{\mu_{ij}}{\sum_{j' \in \mathcal{J}} \mu_{ij'}} : j \in \mathcal{J}\}$, we can simulate $\sum_{j \in \mathcal{J}} \mu_{ij}$ from $Gamma(\sum_{j \in \mathcal{J}} \hat{\alpha_{ij}}, 1)$, where $\hat{\alpha_{ij}}$ are posterior Dirichlet parameters. Eventually we get $\{\mu_{ij} : j \in \mathcal{J}\}$ given $\{\frac{\mu_{ij}}{\sum_{j' \in \mathcal{J}} \mu_{ij'}} : j \in \mathcal{J}\}$ and $\sum_{j \in \mathcal{J}} \mu_{ij}$.

2. $Z_{ij}, \forall ij | \mu_{i'j'}, p_{i'}, \forall i'j', X$
   Let Z_i be a binary vector equals to $(Z_{ij} : j = 1, \ldots, 8)$. Since X is independent across columns (i.e features), We can sample $(Z_i | \mu_{i'j'}, p_{i'}, \forall i'j', X)$ independently across different i's.

$$
\begin{aligned}
&\mathbb{P}(Z_i | \mu_{i'j'}, p_{i'}, \forall i'j', X, Z_{i'}, i' \neq i) \\
&\propto \mathbb{P}(X | \mu_{i'j'}, p_{i'}, \forall i'j', Z_i, Z_{i'}, i' \neq i) \times \mathbb{P}(Z_i | \mu_{i'j'}, p_{i'}, \forall i'j', Z_{i'}, i' \neq i) \\
&\propto \mathbb{P}(X | \mu_{i'j'}, p_{i'}, \forall i'j', Z_i, Z_{i'}, i' \neq i) \\
&\propto \mathbb{P}(X_i | \mu_{ij'}, p_i, \forall j', Z_i)
\end{aligned}
$$

   where X_i indicates ith column of X. Since Z_i can only have 2^8 possible values, it is a discreate distribution and is easy to find out.

3. $p_i, \forall i | Z_{i'j'}, \mu_{ij}, X \sim Beta(\alpha_i + \#(Z_{ij} = 1), \beta_i + \#(Z_{ij} = 0))$
   where α_i, β_i are parameters for prior distribution of p_i.

By sampling μ_{ij}, Z_{ij}, p_i iteratively, we can eventually simulate joint distribution $\mathbb{P}(p_i, Z_{ij}, \mu_{ij} | X)$.