CRAFT: ClusteR-specific Assorted Feature selecTion (Supplementary)

6 Supplementary Material

We now derive the various objectives for the CRAFT framework. We first show the derivation for the generic objective that accomplishes feature selection on the assorted data. We then derive the degenerate cases when all features are retained and all data are (a) numeric, and (b) binary categorical. In particular, when the data are all numeric, we recover the DP-means objective [14].

6.1 Main Derivation: Clustering with Assorted Feature Selection

We have the total number of features, $D = |Cat| + |Num|$. We define $S_{N,k}$ to be the number of points assigned to cluster k. First, note that a Beta distribution with mean c_1 and variance c_2 has shape parameters $\frac{c_1^2(1-c_1)}{c_2} - c_1$ and $\frac{c_1(1-c_1)^2}{c_2} + c_1 - 1$. Therefore, we can find the shape parameters corresponding to m and ρ. Now, recall that for numeric data, we assume the density is of the form $f(x_{nd}|v_{kd})$

$$= \frac{1}{Z_{kd}} e^{-\left[v_{kd}\frac{(x_{nd}-\zeta_{kd})^2}{2\sigma_{kd}^2} + (1-v_{kd})\frac{(x_{nd}-\zeta_d)^2}{2\sigma_d^2}\right]}, \quad (3)$$

where Z_{kd} ensures that the area under the density is 1. Assuming an uninformative conjugate prior on the (numeric) means, i.e. a Gaussian distribution with infinite variance, and using the Iverson bracket notation for discrete (categorical) data, we obtain the joint probability distribution given in Fig. 6 for the underlying graphical model shown in Fig. 1. Note that joint distribution factorizes into a product of posterior distributions (e.g. the beta prior on the features conjugates with the feature likelihood to yield one posterior. We will show that under asymptotic conditions, minimizing the joint negative log-likelihood yields an intuitive objective function via simplification of the log-posteriors.

The total contribution of (3) to the negative joint log-likelihood

$$\begin{aligned} &= \sum_{k=1}^{K^+} \sum_{d \in Num} \sum_{n:z_{n,k}=1} \left[v_{kd}\frac{(x_{nd}-\zeta_{kd})^2}{2\sigma_{kd}^2} \right. && (5) \\ &+ \left. (1-v_{kd})\frac{(x_{nd}-\zeta_d)^2}{2\sigma_d^2} \right] + \sum_{k=1}^{K^+} \sum_{d \in Num} \log Z_{kd}. \end{aligned}$$

The contribution of the selected categorical features depends on the categorical means of the clusters, and is given by

$$-\log\left(\prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \prod_{d \in Cat:v_{kd}=1} \prod_{t \in \mathcal{T}_d} \eta_{kdt}^{\mathbb{I}(x_{nd}=t)} \right).$$

On the other hand, the categorical features not selected are assumed to be drawn from cluster-independent global means, and therefore contribute

$$-\log\left(\prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \prod_{d \in Cat:v_{kd}=0} \prod_{t \in \mathcal{T}_d} \eta_{0dt}^{\mathbb{I}(x_{nd}=t)} \right).$$

Thus, the total contribution of the categorical features is

$$\begin{aligned} -\sum_{k=1}^{K^+} \sum_{n:z_{n,k}=1} &\left[\sum_{d \in Cat:v_{kd}=1} \sum_{t \in \mathcal{T}_d} \mathbb{I}(x_{nd}=t) \log \eta_{kdt} \right. \\ &+ \left. \sum_{d \in Cat:v_{kd}=0} \sum_{t \in \mathcal{T}_d} \mathbb{I}(x_{nd}=t) \log \eta_{0dt} \right]. \end{aligned}$$

The Bernoulli likelihood on v_{kd} couples with the conjugate Beta prior on ν_{kd}. To avoid having to provide the value of ν_{kd} as a parameter, we take its point estimate to be the mean of the resulting Beta posterior, i.e., we set

$$\nu_{kd} = \frac{\left(\frac{m^2(1-m)}{\rho} - m\right) + v_{kd}}{\frac{m(1-m)}{\rho}} = \frac{a_{kd}}{a_{kd}+b_{kd}}, \quad (6)$$

where

$$\begin{aligned} a_{kd} &= \frac{m^2(1-m)}{\rho} - m + v_{kd}, \text{ and} \\ b_{kd} &= \frac{m(1-m)^2}{\rho} + m - v_{kd}. \end{aligned}$$

Then the contribution of the posterior to the negative log likelihood is

$$-\sum_{k=1}^{K^+} \sum_{d=1}^{D} \left[\log\left(\frac{a_{kd}}{a_{kd}+b_{kd}}\right)^{a_{kd}} + \log\left(\frac{b_{kd}}{a_{kd}+b_{kd}}\right)^{b_{kd}} \right],$$

or equivalently,

$$\sum_{k=1}^{K^+} \sum_{d=1}^{D} \underbrace{\left[\log\left(a_{kd}+b_{kd}\right)^{(a_{kd}+b_{kd})} - \log a_{kd}^{a_{kd}} - \log b_{kd}^{b_{kd}} \right]}_{F(v_{kd})}.$$

$$
\begin{aligned}
&\mathbb{P}(x, z, v, \nu, \eta, \zeta, m) \\
&= \mathbb{P}(x|z, v, \eta, \zeta)\mathbb{P}(v|\nu)\mathbb{P}(z)\mathbb{P}(\eta)\mathbb{P}(\nu; m, \rho) \\
&= \prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \left[\left(\prod_{d \in Cat: v_{kd}=1} \prod_{t \in \mathcal{T}_d} \eta_{kdt}^{\mathbb{I}(x_{nd}=t)} \right) \left(\prod_{d \in Cat: v_{kd}=0} \prod_{t \in \mathcal{T}_d} \eta_{0dt}^{\mathbb{I}(x_{nd}=t)} \right) \right. \\
&\quad \left. \left(\prod_{d' \in Num} \frac{1}{Z_{kd'}} e^{-\left[v_{kd'}(x_{nd'}-\zeta_{kd'})^2/(2\sigma_{kd'}^2)+(1-v_{kd'})(x_{nd'}-\zeta_{d'})^2/(2\sigma_{d'}^2))\right]} \right) \right] \\
&\quad \cdot \left[\prod_{k=1}^{K^+} \prod_{d=1}^{D} \nu_{kd}^{v_{kd}} (1-\nu_{kd})^{1-v_{kd}} \right] \cdot \left[\theta^{K^+-1} \frac{\Gamma\left(\theta+1\right)}{\Gamma\left(\theta+N\right)} \prod_{k=1}^{K^+} (S_{N,k}-1)! \right] \\
&\quad \cdot \left[\prod_{k=1}^{K^+} \prod_{d \in Cat} \frac{\Gamma\left(\sum_{t \in \mathcal{T}_d} \dfrac{\alpha_{kdt}}{K^+}\right)}{\prod_{t \in \mathcal{T}_d} \Gamma\left(\dfrac{\alpha_{kdt}}{K^+}\right)} \prod_{t' \in \mathcal{T}_d} \eta_{kdt'}^{(\alpha_{kdt'}/K^+)-1} \right] \\
&\quad \cdot \prod_{k=1}^{K^+} \prod_{d=1}^{D} \frac{\Gamma\left(\dfrac{m(1-m)}{\rho}-1\right) \nu_{kd}^{\left(\frac{m^2(1-m)}{\rho}-m-1\right)} (1-\nu_{kd})^{\left(\frac{m(1-m)^2}{\rho}-(2-m)\right)}}{\Gamma\left(\dfrac{m^2(1-m)}{\rho}-m\right)\Gamma\left(\dfrac{m(1-m)^2}{\rho}-(1-m)\right)}
\end{aligned} \tag{4}
$$

Figure 6: Joint probability distribution for the generic case (both numeric and categorical features).

Since $v_{kd} \in \{0, 1\}$, this simplifies to

$$
\begin{aligned}
\sum_{k=1}^{K^+}\sum_{d=1}^{D} F(v_{kd}) &= \sum_{k=1}^{K^+}\sum_{d=1}^{D} \left[v_{kd}(F(1)-F(0)) + F(0)\right] \\
&= \left(\sum_{k=1}^{K^+}\sum_{d=1}^{D} v_{kd} \right) \Delta F + K^+ D F(0),
\end{aligned} \tag{7}
$$

where $\Delta F = F(1) - F(0)$ quantifies the change when a feature is selected for a cluster.

The numeric means do not make any contribution since we assumed an uninformative conjugate prior over $\mathbb{R}$. On the other hand, the categorical means contribute

$$
-\log \left[\prod_{k=1}^{K^+} \prod_{d \in Cat} \frac{\Gamma\left(\sum_{t \in \mathcal{T}_d} \dfrac{\alpha_{kdt}}{K^+}\right)}{\prod_{t \in \mathcal{T}_d} \Gamma\left(\dfrac{\alpha_{kdt}}{K^+}\right)} \prod_{t' \in \mathcal{T}_d} \eta_{kdt'}^{(\alpha_{kdt'}/K^+)-1} \right],
$$

which simplifies to

$$
\begin{aligned}
\sum_{k=1}^{K^+} \sum_{d \in Cat} \Bigg[&-\log \frac{\Gamma\left(\sum_{t \in \mathcal{T}_d} \dfrac{\alpha_{kdt}}{K^+}\right)}{\prod_{t \in \mathcal{T}_d} \Gamma\left(\dfrac{\alpha_{kdt}}{K^+}\right)} \\
&- \sum_{t' \in \mathcal{T}_d} \left(\frac{\alpha_{kdt'}}{K^+} - 1 \right) \log \eta_{kdt'} \Bigg].
\end{aligned} \tag{8}
$$

Finally, the Dirichlet process specifies a distribution over possible clusterings, while favoring assignments of points to a small number of clusters. The contribution of the corresponding term is

$$
-\log \left[\theta^{K^+-1} \frac{\Gamma\left(\theta+1\right)}{\Gamma\left(\theta+N\right)} \prod_{k=1}^{K^+} (S_{N,k}-1)! \right],
$$

or equivalently,

$$
-(K^+-1)\log\theta - \log \left(\frac{\Gamma\left(\theta+1\right)}{\Gamma\left(\theta+N\right)} \prod_{k=1}^{K^+} (S_{N,k}-1)! \right). \tag{9}
$$

The total negative log-likelihood is just the sum of terms in (5), (6), (7), (8), and (9). We want to maximize the joint likelihood, or equivalently, minimize the total negative log-likelihood. We would use asymptotics to simplify our objective. In particular, letting $\sigma_d \to \infty,\ \forall k \in [K^+]$ and $d \in Num$, and $\alpha_{kdt} \to K^+,\ \forall t \in \mathcal{T}_d,\ d \in Cat, k \in [K^+]$, setting $\log\theta$ to

$$
-\left(\lambda + \frac{\displaystyle\sum_{k=1}^{K^+}\sum_{d \in Cat} \log|\mathcal{T}_d| - \sum_{k=1}^{K^+}\sum_{d \in Num} \log Z_{kd}}{K^+-1} \right),
$$

and ignoring the term containing $S_{N,k}$ that contributes $\mathcal{O}(1)$, we obtain our objective for assorted feature selection:

$$\underset{z,v,\eta,\zeta,\sigma}{\arg\min}\ \underbrace{\sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{d\in Num}\frac{v_{kd}(x_{nd}-\zeta_{kd})^2}{2\sigma_{kd}^2}}_{\text{Numeric Data Discrepancy}}$$

$$+\underbrace{(\lambda+DF_0)K^+}_{\text{Regularization Term}}+\underbrace{\left(\sum_{k=1}^{K^+}\sum_{d=1}^{D}v_{kd}\right)F_\Delta}_{\text{Feature Control}}$$

$$+\underbrace{\sum_{k=1}^{K^+}\sum_{d\in Cat}\left[v_{kd}\left(\sum_{n:z_{n,k}=1}-\mathbb{I}(x_{nd}=t)\log\eta_{kdt}\right)\right.}_{\text{Categorical Discrepancy Term I}}$$

$$\underbrace{\left.+(1-v_{kd})\sum_{n:z_{n,k}=1}\sum_{t\in\mathcal{T}_d}-\mathbb{I}(x_{nd}=t)\log\eta_{0dt}\right]}_{\text{Categorical Discrepancy Term II}},$$

where $\Delta F = F(1)-F(0)$ quantifies the change when a feature is selected for a cluster, and we have renamed the constants $F(0)$ and ΔF as F_0 and F_Δ respectively.

6.1.1 Setting ρ

Reproducing the equation for ν_{kd} from (6), since we want to ensure that $\nu_{kd}\in(0,1)$, we must have

$$0<\frac{\left(\dfrac{m^2(1-m)}{\rho}-m\right)+v_{kd}}{\dfrac{m(1-m)}{\rho}}<1.$$

Since $v_{kd}\in\{0,1\}$, this immediately constrains

$$\rho\in(0,m(1-m)).$$

Note that ρ guides the selection of features: a high value of ρ, close to $m(1-m)$, enables local feature selection (v_{kd} becomes important), whereas a low value of ρ, close to 0, reduces the influence of v_{kd} considerably, thereby resulting in global selection.

6.2 Degenerate Case: Clustering Binary Categorical Data without Feature Selection

In this case, the discrete distribution degenerates to Bernoulli, while the numeric discrepancy and the feature control terms do not arise. Therefore, we can replace the Iverson bracket notation by having cluster means μ drawn from Bernoulli distributions. Then, the joint distribution of the observed data x, cluster indicators z and cluster means μ is given by $\mathbb{P}(x,z,\mu)$

$$\begin{aligned}
&= \mathbb{P}(x|z,\mu)\mathbb{P}(z)\mathbb{P}(\mu)\\
&= \underbrace{\left[\prod_{k=1}^{K^+}\prod_{n:z_{n,k}=1}\prod_{d=1}^{D}\mu_{kd}^{x_{nd}}(1-\mu_{kd})^{1-x_{nd}}\right]}_{(A)}\\
&\cdot\ \underbrace{\left[\theta^{K^+-1}\frac{\Gamma(\theta+1)}{\Gamma(\theta+N)}\prod_{k=1}^{K^+}(S_{N,k}-1)!\right]}_{(B)} \qquad (10)\\
&\cdot\underbrace{\left[\prod_{k=1}^{K^+}\prod_{d=1}^{D}\frac{\Gamma\left(\dfrac{\alpha}{K^+}+1\right)}{\Gamma\left(\dfrac{\alpha}{K^+}\right)\Gamma(1)}\mu_{kd}^{\frac{\alpha}{K^+}-1}(1-\mu_{kd})^0\right]}_{(C)}.
\end{aligned}$$

The joint negative log-likelihood is

$$-\log\mathbb{P}(x,z,\mu)=-[\log(A)+\log(B)+\log(C)].$$

We first note that $\log(A)$

$$\begin{aligned}
&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{d=1}^{D}x_{nd}\log\mu_{kd}+(1-x_{nd})\log(1-\mu_{kd})\\
&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{d=1}^{D}x_{nd}\log\left(\frac{\mu_{kd}}{1-\mu_{kd}}\right)+\log(1-\mu_{kd})\\
&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{d=1}^{D}\left[\log(1-\mu_{kd})+\mu_{kd}\log\left(\frac{\mu_{kd}}{1-\mu_{kd}}\right)\right.\\
&\quad\left.+\ x_{nd}\log\left(\frac{\mu_{kd}}{1-\mu_{kd}}\right)-\mu_{kd}\log\left(\frac{\mu_{kd}}{1-\mu_{kd}}\right)\right]\\
&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{d=1}^{D}\left[(x_{nd}-\mu_{kd})\log\left(\frac{\mu_{kd}}{1-\mu_{kd}}\right)\right.\\
&\quad\left.+\ \mu_{kd}\log\mu_{kd}+(1-\mu_{kd})\log(1-\mu_{kd})\right]\\
&= \sum_{k=1}^{K^+}\sum_{n:z_{n,k}=1}\sum_{d=1}^{D}(x_{nd}-\mu_{kd})\log\left(\frac{\mu_{kd}}{1-\mu_{kd}}\right)-\mathbb{H}(\mu_{kd}),
\end{aligned}$$

where

$$\mathbb{H}(p)=-p\log p-(1-p)\log(1-p)\ \text{ for } p\in[0,1].$$

$\log(B)$ and $\log(C)$ can be computed via steps analogous to those used in assorted feature selection. Invoking the asymptotics by letting $\alpha\to K^+$, setting

$$\theta=e^{-\left(\lambda+\dfrac{K^+D}{K^+-1}\log\left(\dfrac{\alpha}{K^+}\right)\right)},$$

and ignoring the term containing $S_{N,k}$ that contributes $\mathcal{O}(1)$, we obtain the following objective:

$$\operatorname*{argmin}_{z,\mu} \sum_{k=1}^{K^+} \lambda K^+ + \sum_{n:z_{n,k}=1} \sum_{d} \underbrace{\left[\mathbb{H}(\mu_{kd}) + (\mu_{kd} - x_{nd}) \log \left(\frac{\mu_{kd}}{1-\mu_{kd}} \right) \right]}_{\text{(Binary Discrepancy)}},$$

where the term (Binary Discrepancy) is an objective for binary categorical data, similar to the K-means objective for numeric data. This suggests a very intuitive procedure, which is outlined in Algorithm 3.

Algorithm 3 Clustering binary categorical data

```
Input: $x_1, \ldots, x_N \in \{0,1\}^D$: binary categorical data, and $\lambda > 0$: cluster penalty parameter.
Output: $K^+$: number of clusters and $l_1, \ldots, l_{K^+}$: clustering.
  1. Initialize $K^+ = 1$, $l_1 = \{x_1, \ldots, x_N\}$ and the mean $\mu_1$ (sample randomly from the dataset).
  2. Initialize cluster indicators $z_n = 1$ for all $n \in [N]$.
  3. Repeat until convergence
     • Compute $\forall k \in [K^+], d \in [D]$:
           $\mathbb{H}(\mu_{kd}) = -\mu_{kd} \log \mu_{kd} - (1-\mu_{kd}) \log(1-\mu_{kd})$.
     • For each point $x_n$
         – Compute the following for all $k \in [K^+]$:
             $d_{nk} = \sum_{d=1}^{D} \left[ \mathbb{H}(\mu_{kd}) + (\mu_{kd} - x_{nd}) \log \left( \frac{\mu_{kd}}{1-\mu_{kd}} \right) \right]$.
         – If $\min_k d_{nk} > \lambda$, set $K^+ = K^+ + 1$, $z_n = K^+$, and $\mu_{K^+} = x_n$.
         – Otherwise, set $z_n = \operatorname*{argmin}_k d_{nk}$.
     • Generate clusters $l_1, \ldots, l_{K^+}$ based on $z_1, \ldots, z_{K^+}$: $l_k = \{x_n \mid z_n = k\}$.
     • For each cluster $l_k$, update $\mu_k = \frac{1}{|l_k|} \sum_{x \in l_k} x$.
```

In each iteration, the algorithm computes "distances" to the cluster means for each point to the existing cluster centers, and checks if the minimum distance is within λ. If yes, the point is assigned to the nearest cluster, otherwise a new cluster is started with the point as its cluster center. The cluster means are updated at the end of each iteration, and the steps are repeated until there is no change in cluster assignments over successive iterations.

We get a more intuitively appealing objective by noting that the objective (11) can be equivalently written as

$$\operatorname*{argmin}_{z} \sum_{k=1}^{K^+} \sum_{n:z_{n,k}=1} \sum_{d} \mathbb{H}(\mu^*_{kd}) + \lambda K^+, \qquad (11)$$

where μ^*_{kd} denotes the mean of feature d computed by using points belonging to cluster k. characterizes the uncertainty. Thus the objective tries to minimize the overall uncertainty across clusters and thus forces similar points to come together. The regularization term ensures that the points do not form too many clusters, since in the absence of the regularizer each point will form a singleton cluster thereby leading to a trivial clustering.

6.3 Degenerate Case: Clustering Numerical Data without Feature Selection (Recovering DP-means)

In this case, there are no categorical terms. Furthermore, assuming an uninformative conjugate prior on the numeric means, the terms that contribute to the negative joint log-likelihood are

$$\prod_{k=1}^{K^+} \prod_{d'} \frac{1}{Z_{kd'}} e^{-\left[\frac{v_{kd'}(x_{nd'} - \zeta_{kd'})^2}{(2\sigma^2_{kd'})} + (1-v_{kd'}) \frac{(x_{nd'} - \zeta_{d'})^2}{(2\sigma^2_{d'})} \right]},$$

and

$$\theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k} - 1)!.$$

Taking the negative logarithms on both these terms and adding them up, setting $\log \theta$ to

$$-\left(\lambda + \frac{\sum_{k=1}^{K^+} \sum_{d'} \log Z_{kd'}}{K^+ - 1} \right),$$

and $v_{kd'} = 1$ (since all features are retained), letting $\sigma_{d'} \to \infty$ for all d', and ignoring the $\mathcal{O}(1)$ term containing $S_{N,k}$, we obtain

$$\operatorname*{argmin}_{z} \sum_{k=1}^{K^+} \sum_{n:z_{n,k}=1} \sum_{d} \frac{(x_{nd} - \zeta^*_{kd})^2}{2\sigma^{*2}_{kd}} + \lambda K^+, \qquad (12)$$

where ζ^*_{kd} and σ^{*2}_{kd} are, respectively, the mean and variance of the feature d computed using all the points assigned to cluster k. This degenerates to the DP-means objective [14] when $\sigma^*_{kd} = 1/\sqrt{2}$, for all k and d. Thus, using a completely different model and analysis to [14], we recover the DP-means objective as a special case.