Kernel Means Matching: Cross Entropy

Madison Cooley

August 20, 2018

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{C} y_{nj} log(\hat{y}_{nj})$$

$$\hat{y}_{nj} = \frac{e^{s_{nj}}}{\sum_{c=1}^{C} e^{s_{nc}}}$$

$$s_{nj} = B_j.Ax_n$$

$A \in \mathbb{R}^{D\times P}$
$B \in \mathbb{R}^{C\times D}$
$x_n \in \mathbb{R}^{P\times 1}$

Adding the re-weighting coefficient (β_n) from "Correcting Sample Selection Bias by Unlabeled Data" and a regularization term gives...

$$L = -\sum_{n=1}^{N}\sum_{j=1}^{C} \beta_n y_{nj} log(\hat{y}_{nj}) + \frac{\lambda}{2}||\theta||^2$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{C} -\beta_n y_{nj} log(p(\hat{y}_{nj}|\phi(x_n);\theta)) + \frac{\lambda}{2}||\theta||^2$$

Using the exponential families approach where,

$$log(p(y|x;\theta)) = \langle\phi(x,y),\theta\rangle - g(\theta|x)$$

where,

$$g(\theta|x) = log\sum_{y\in Y} exp(\langle\phi(x,y),\theta\rangle)$$

$$L = \sum_{n=1}^{N}\sum_{j=1}^{C} -\beta_n y_{nj}[\langle\phi(x_n,\hat{y}_{nj}),\theta\rangle - log\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}} exp(\langle\phi(x_n,\hat{y}_{nj}),\theta\rangle)] + \frac{\lambda}{2}||\theta||^2$$

Now, finding the derivative w.r.t. θ,

$$\frac{dL}{d\theta} = \sum_{n=1}^{N}\sum_{j=1}^{C} -\beta_n y_{nj}\phi(x_n,\hat{y}_{nj}) + \beta_n y_{nj}\frac{\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}} exp(\langle\phi(x_n,\hat{y}_{nj}),\theta\rangle)}{\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}} ln(10)exp(\langle\phi(x_n,\hat{y}_{nj}),\theta\rangle)}\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}}\phi(x_n,\hat{y}_{nj}) - \lambda\theta$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{C} -\beta_n y_{nj}\phi(x_n,\hat{y}_{nj}) + \beta_n y_{nj}\frac{1}{ln(10)}\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}}\phi(x_n,\hat{y}_{nj}) - \lambda\theta = 0$$

Solve for θ,

$$= \sum_{n=1}^{N}\sum_{j=1}^{C} -\beta_n y_{nj}\phi(x_n,\hat{y}_{nj}) + \beta_n y_{nj}\frac{1}{ln(10)}\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}}\phi(x_n,\hat{y}_{nj}) = \lambda\theta$$

$$\theta = \sum_{n=1}^{N}\sum_{j=1}^{C}\alpha_j\phi(x_n,\hat{y}_{nj})$$

Where $\alpha_{j'} = (\frac{1}{\lambda})[\beta_n y_{nj}\frac{1}{ln(10)} + \beta_n y_{nj}]$

Substituting theta back in to L gives...

$$L = \sum_{n,n'=1}^{N}\sum_{j,j'=1}^{C}\beta_n y_{nj} log\sum_{\hat{y}_{nj}\in\hat{Y}_{nj}} exp(\langle\phi(x_n,\hat{y}_{nj}),\alpha_{j'}\phi(x_{n'},\hat{y}_{nj'})\rangle)$$

$$-\beta_n y_{nj}\langle\phi(x_n,\hat{y}_{nj}),\alpha_{j'}\phi(x_{n'},\hat{y}_{nj'})\rangle + \frac{\lambda}{2}||\alpha_{j'}\phi(x_{n'},\hat{y}_{nj'})||^2$$

$$= \sum_{n,n'=1}^{N}\sum_{j,j'=1}^{C}\beta_n y_{nj} g(\alpha|x_{n'}) - \beta_n y_{nj}\alpha_{j'}\phi(x_{n'},\hat{y}_{nj'})\phi(x_n,\hat{y}_{nj}) + \frac{\lambda}{2}\alpha_{j'}\alpha_{j'}\phi(x_{n'},\hat{y}_{nj'})\phi(x_{n'},\hat{y}_{nj'}),$$

where $g(\alpha|x_{n'}) = log\sum_{\hat{y}_{nj'}\in\hat{Y}_{nj'}} exp(\alpha_{j'}\phi(x_{n'},\hat{y}_{nj'})\phi(x_n,\hat{y}_{nj}))$

$$\boxed{minimize \sum_{n,n'=1}^{N}\sum_{j,j'=1}^{C}\beta_n y_{nj} g(\alpha|x_{n'}) - \beta_n y_{nj} k(x_{n'},\hat{y}_{nj'},x_n,\hat{y}_{nj}) + \tfrac{\lambda}{2}\alpha_{j'}\alpha_{j'}k(x_{n'},\hat{y}_{nj'},x_{n'},\hat{y}_{nj'})}$$

where $g(\alpha|x_{n'}) = log\sum_{\hat{y}_{nj'}\in\hat{Y}_{nj'}} exp(\alpha_{j'}k(x_{n'},\hat{y}_{nj'},x_{n'},\hat{y}_{nj'}))$

Explanation:

One method to correct sample selection bias is called Kernel Mean Matching or KMM, which was proposed in "Correcting Sample Selection Bias by Unlabeled Data." This method re-weights training points such that the means of the training and testing points are close in a reproducing kernel Hilbert space (RKHS). The goal is to basically re-weight the training data to more closely resemble to testing data.

Optimal β (re-weighting term) is found by...

$$\text{Using } K_{ij} = k(x_i^{tr}, x_j^{tr}) \text{ and } k_i = \frac{n_{tr}}{n_{te}} \sum_{j=1}^{n_{te}} k(x_i^{tr}, x_j^{te})$$

$$||\frac{1}{n^{tr}} \sum_{i=1}^{n^{tr}} \beta_i \Phi(x_i^{tr}) - \frac{1}{n^{te}} \sum_{i=1}^{n^{te}} \beta_i \Phi(x_i^{te})||^2$$

$$= \frac{1}{n_{tr}^2} \beta^T K \beta - \frac{2}{n_{tr}^2} k^T \beta + const$$

$$= minimize_\beta \frac{1}{2} \beta^T K \beta - k^T \beta$$

s.t. $\beta_i \in [0, B]$ and $|\sum_{i=1}^{n_{tr}} \beta_i - n_{tr}| \leq n_{tr}\epsilon$