Text classification with discriminative models

- Discriminative modeling approach that learns P(y|x)

- Quick refresher about MLE & discriminative models:

$$P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta}) \quad \textit{where} \quad \mathbf{x},\mathbf{y} \sim P_D\,, \quad \mathbf{y} \in \{0,1\}^K \longrightarrow$$ one hot encoding

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{M} -\log P(\mathbf{y}_i\,|\,\mathbf{x}_i;\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim P_D}\big[\log P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\big] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; -\sum_{i=1}^{M} \mathbf{y}\log P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta}) \longrightarrow
\end{aligned}$$

Cross entropy between data distribution and the model output distribution, $H\big(P_D(\mathbf{y}\,|\,\mathbf{x}), P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\big)$. Note, this holds for arbitrary PMFs (softmax, Bernoulli, etc...). Because the labels, $\mathbf{y}$, are one hot, they have zero entropy, and thus in this setting minimizing the cross entropy is equivalent to minimizing the KL divergence, $D_{KL}\big(P_D(\mathbf{y}\,|\,\mathbf{x}), P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\big)$.

Generalized softmax regression

- Generic softmax regression adopts the underlying model mapping y <- x in logistic regression, but extends it to >2 dimensions

- The softmax function is more broadly useful for other (not necessarily linear) mapping functions, too! It comes in very handy for any function that's at least once differentiable w.r.t. parameters that are being learned. Here we'll just say there is some function $f : \mathbf{x} \rightarrow \mathbf{y}$ that's differentiable with respect to some parameters, $\boldsymbol{\theta}$. Note, below represents a vector of probabilities, one for each class.

$$P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \frac{e^{f(\mathbf{x};\boldsymbol{\theta})}}{\sum_i e^{f^{(i)}(\mathbf{x};\boldsymbol{\theta})}}$$