

NN parameter estimation for classification

- Model: $P(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta})$ where $\mathbf{x}, \mathbf{y} \sim P_D$, $\mathbf{y} \in \{0,1\}^K$
- Optimization:
$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^M -\mathbf{y}_i \log P(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}) \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} -\mathbb{E}_{\mathbf{x}, \mathbf{y} \sim P_D} [\log P(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta})] \\ &= \operatorname{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}, \mathbf{y} \sim P_D} \left[\log \frac{1}{P(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta})} \right]\end{aligned}$$

one hot encoding



Cross entropy between data distribution and the model output distribution, $H(P_D(\mathbf{y} | \mathbf{x}), P(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}))$. 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, $D_{KL}(P_D(\mathbf{y} | \mathbf{x}), P(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}))$.

