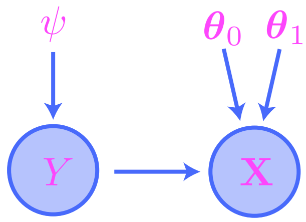
Problem 1: Consider a *Naive Bayes model* as described in the programming assignment for chapter 12. The underlying graph is of the form



where $\mathbf{X} \in \mathbb{R}^n$. The parameters are given by a number $\psi \in [0,1]$ which parametrizes the distribution of $Y \sim \mathcal{B}er(\psi)$, as well as two vectors $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in [0,1]^n$. The link function at $\mathbf{X}$ is given by

$$p(\mathbf{x} \mid y;\ \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \prod_{j=1}^{n} \phi_j^{x_j}(1-\phi_j)^{1-x_j}$$

where

$$\boldsymbol{\phi} = (1-y)\boldsymbol{\theta}_0 + y\boldsymbol{\theta}_1$$

and $\boldsymbol{\phi}^\intercal = (\phi_1, \ldots, \phi_n)$.

(a) Assuming that Naive Bayes models are trained as **generative** models, write down a formula for the model likelihood function $\mathcal{L}_{\text{model}}(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. For simplicity, your formula should contain the ϕ_j's rather than the parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ themselvs.

We have

$$\begin{aligned}\mathcal{L}_{\text{model}}(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1;\ \mathbf{x}, y) &= p(\mathbf{x}, y;\ \psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\\ &= p(y;\psi)p(\mathbf{x} \mid y;\ \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)\\ &= \psi^y(1-\psi)^{1-y}\prod_{j=1}^{n} \phi_j^{x_j}(1-\phi_j)^{1-x_j}.\end{aligned}$$

(b) Using your answer from part (a), write down a formula for the model surprisal function $\mathcal{I}_{\text{model}}(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. For simplicity, your formula should contain the ϕ_j's rather than the parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ themselves.

We have

$$\begin{aligned}\mathcal{I}_{\text{model}}(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1;\ \mathbf{x}, y) &= -\log \mathcal{L}_{\text{model}}(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1;\ \mathbf{x}, y)\\ &= -y\log\psi - (1-y)\log(1-\psi) - \sum_{j=1}^{n}\left[x_j \log\phi_j + (1-x_j)\log(1-\phi_j)\right].\end{aligned}$$

(c) Using your answer from part (b), write down an explicit formula for the cross entropy stochastic objective function $J(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ for a dataset of size m.

Letting

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m) \in \{0,1\}^m \times \{0,1\}$$

be the dataset, by definition we have

$$\begin{aligned}
J(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= \frac{1}{m}\sum_{i=1}^{m} \mathcal{I}_{\text{model}}(\psi, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1;\ \mathbf{x}_i, y_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} \left\{ -y_i \log \psi - (1-y_i)\log(1-\psi) - \sum_{j=1}^{n} \left[x_{ij}\log \phi_j + (1-x_{ij})\log(1-\phi_j) \right] \right\}.
\end{aligned}$$