MAT 180 Theoretical Homework 3

Problem 1: Consider the Naive Bayes classification problem in the case that x_i takes values in $\{1..., r\}$ and y takes values in $\{1, ..., s\}$ for some $r, s \in \mathbb{N}$. Use multinoulli (categorical) model distributions for both $p_{\text{data}}(\mathrm{y})$ and $p_{\text{data}}(\mathrm{x}_j|\mathrm{y})$. Compute the maximum likelihood values for parameters in both models using a dataset $\mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{y} \in \mathbb{R}^m$ assumed to be I.I.D and sampled from p_{data}. You should derive your formulas from the equations

$$p_{\text{model}}(\mathrm{y} = l; \boldsymbol{\theta}) = \boldsymbol{\theta}_l \qquad \boldsymbol{\theta} \in \mathbb{R}^s \qquad \sum_{l=1}^{s} \boldsymbol{\theta}_l = 1$$

$$\boldsymbol{\theta}_{\text{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{m} \log(p_{\text{model}}(\mathrm{y} = \mathbf{y}_i; \boldsymbol{\theta}))$$

$$p_{\text{model}}(\mathrm{x}_j = k | \mathrm{y} = l; \boldsymbol{\theta}^{(j)}) = \boldsymbol{\theta}^{(j)}_{(l,k)} \qquad \boldsymbol{\theta}^{(j)} \in \mathbb{R}^{s \times r} \qquad \forall l, \ \sum_{k=1}^{r} \boldsymbol{\theta}^{(j)}_{(l,k)} = 1$$

$$\boldsymbol{\theta}^{(j)}_{\text{ML}} = \underset{\boldsymbol{\theta}^{(j)}}{\operatorname{argmax}} \sum_{i=1}^{m} \log(p_{\text{model}}(\mathrm{x} = X_{i,j} | \mathrm{y} = \mathbf{y}_i; \boldsymbol{\theta}^{(j)}))$$

Hint: this can be viewed as a constrained optimization problem (with no inequalities... so a Lagrange multipliers problem). Use the Karush-Kuhn-Tucker conditions to optimize. For $\boldsymbol{\theta}^{(j)}$ you will need to take the gradient with respect to a matrix rather than a vector. You can just treat the matrix as a vector and take the gradient as you would with a vector, except instead of organizing the result as a vector, you organize it as a matrix. For example, if $\boldsymbol{\theta} \in \mathbb{R}^{2\times 2}$ and $f(\boldsymbol{\theta}) = \boldsymbol{\theta}^2_{(1,1)} + 2\boldsymbol{\theta}_{(1,2)}\boldsymbol{\theta}_{(2,1)} - 2\boldsymbol{\theta}^2_{(2,2)}$, then

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \begin{pmatrix} 2\boldsymbol{\theta}_{(1,1)} & 2\boldsymbol{\theta}_{(2,1)} \\ 2\boldsymbol{\theta}_{(1,2)} & -4\boldsymbol{\theta}_{(2,2)} \end{pmatrix}$$

Problem 2: Consider the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ with SVD $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ and suppose each column of X has mean 0.

1. Show that for any vector $\mathbf{v} \in \mathbb{R}^n$, $\text{Mean}(\mathbf{Xv}) = 0$ and $\text{Var}(\mathbf{Xv}) = \frac{\mathbf{v}^\top \mathbf{X}^\top \mathbf{X} \mathbf{v}}{m}$
2. Show that for any eigenvectors $\mathbf{v}_i, \mathbf{v}_j$ of $\mathbf{X}^\top\mathbf{X}$ with $i \neq j$ and any $a, b \in \mathbb{R}$ we have $\text{Var}(a\mathbf{X}\mathbf{v}_i + b\mathbf{X}\mathbf{v}_j) = \frac{a^2\sigma_i^2}{m} + \frac{b^2\sigma_j^2}{m}$
3. Prove that for $1 \leq i \leq n$ we have that

$$\mathbf{v}_i = \underset{\substack{||\mathbf{v}||_2 = 1 \\ \mathbf{v} \notin \langle \mathbf{v}_1, \ldots, \mathbf{v}_{i-1} \rangle}}{\operatorname{argmax}} \ \text{Var}(\mathbf{Xv})$$

where $\langle \mathbf{v}_1, \ldots, \mathbf{v}_{i-1} \rangle$ denotes the span of the first i column vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$ of V. If desired, you can use the following fact about convex combinations of real numbers (but you should prove it!): let $x_1, \ldots, x_n \in \mathbb{R}$ and $a_1, \ldots, a_n \geq 0$ such that $\sum_i a_i = 1$. Then

$$\min_i(x_i) \leq \sum_i a_i x_i \leq \max_i(x_i).$$

4. Interpret this as follows. Viewing $\mathbf{X} \in \mathbb{R}^{m \times n}$ as a data matrix, giving m points in $\mathbb{R}^n$ (one for each row), the PCA represents $\mathbf{X}$ as $\mathbf{Z} = \mathbf{X}\mathbf{V}_k$ for some $1 \leq k \leq n$. For $1 \leq j \leq n$ define $\mathrm{Proj}_{\mathbf{v}_j}(\mathbf{X}) \in \mathbb{R}^m$ so that

$$\mathrm{Proj}_{\mathbf{v}_j}(\mathbf{X})_i = \mathbf{v}_j^\top \mathrm{Proj}_{\mathbf{v}_j}(\mathbf{X}_{i,*})$$

so the ith entry of $\mathrm{Proj}_{\mathbf{v}_j}(\mathbf{X})$ is the *scalar projection* of the ith data point onto the direction vector $\mathbf{v}_j$. Show that the vectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ give an orthogonal basis of $\mathbb{R}^n$ with the property that

$$\mathrm{Var}(\mathrm{Proj}_{\mathbf{v}_1}(\mathbf{X})) \geq \mathrm{Var}(\mathrm{Proj}_{\mathbf{v}_2}(\mathbf{X})) \geq \cdots \geq \mathrm{Var}(\mathrm{Proj}_{\mathbf{v}_n}(\mathbf{X})) \geq 0.$$

5. Based on this interpretation, write down the expected matrices Σ, V for the SVD for a data matrix of points in $\mathbb{R}^3$ generated by the normal distributions

$$\begin{aligned} p_{\mathrm{data}}(x) &= \mathcal{N}(x; 0; 5) \\ p_{\mathrm{data}}(y) &= \mathcal{N}(y; 0; 2) \\ p_{\mathrm{data}}(z) &= \mathcal{N}(z; 0; 10) \end{aligned}$$

Hint: you can check your answer using Python. Numpy has the method

```
np.random.normal(m,s)
```

to sample a normal distribution with mean m and variance s^2.