156 Assignment 5 solutions

Problem 1: Show that the derivative of the error function

$$E = -\sum_{k=1}^{K} t_k \log(y_k)$$

with respect to the activation a_k for output units having the softmax activation function

$$y_k(a) = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}}$$

satisfies

$$\frac{\partial E}{\partial a_k} = y_k - t_k.$$

Solution. Distributing across the sum,

$$\frac{\partial E}{\partial a_k} = -\sum_{j=1}^{K} \frac{t_j}{y_j}\frac{\partial y_j}{\partial a_k}.$$

Note that, when $j \neq k$,

$$\frac{\partial y_j}{\partial a_k} = -\frac{e^{a_j} e^{a_k}}{(\sum_{j=1}^{K} e^{a_j})^2} = -y_j y_k;$$

also, when $j = k$,

$$\frac{\partial y_j}{\partial a_k} = -\frac{e^{a_k} e^{a_k}}{(\sum_{j=1}^{K} e^{a_j})^2} + \frac{e^{a_k}}{(\sum_{j=1}^{K} e^{a_j})^2} = -y_k^2 + y_k.$$

Thus,

$$\frac{\partial E}{\partial a_k} = \left[-\sum_{j=1}^{K} \frac{t_j}{y_j}(-y_j y_k)\right] - \frac{t_k}{y_k} y_k = \left[\sum_{j=1}^{K} t_j\right] y_k - t_k.$$

Finally, by virtue of the 1-in-K encoding, we have $\sum_{j=1}^{K} t_j = 1$; hence, we obtain the desired

$$\frac{\partial E}{\partial a_k} = y_k - t_k.$$

$\square$

Problem 3: We show by induction that the linear projection onto an M-dimensional subspace that maximizes the variance of the projected data is defined by the M eigenvectors of the data covariance matrix $\mathbf{S}$, given by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^\top.$$

In Section 12.1 of the textbook, this was proven in the special case $M = 1$. We wish to use induction to prove the case for general M. To this end, assume the result holds for some particular M, and show the result for projections onto subspaces of dimension $M + 1$. Use the following approach:

- Set up a Lagrange multiplier formulation of the constrained optimization problem of maximizing the variance of the projected data, subject to the constraints of orthogonality and unit length.
- Use orthonormality to show that the solution vector $\mathbf{u}_{M+1}$ is also an eigenvector of $\mathbf{S}$.
- Show that the variance is maximized in case that $\mathbf{u}_{M+1}$ is chosen to correspond to the $(M+1)$-st largest eigenvalue.

Solution. As indicated, we accept the result for $M = 1$; by induction, we assume that the result holds also for M, and consider the case $M + 1$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_M$ be the eigenvectors corresponding to the M largest eigenvalues. Let $\mathbf{u}_{M+1}$ be a unit vector which is orthogonal to each of the $\mathbf{u}_j$, $1 \leq j \leq M$. Then the orthogonal projection of a vector $\mathbf{y}$ onto the subspace spanned by $\mathbf{u}_1, \ldots, \mathbf{u}_{M+1}$ takes the form

$$\sum_{j=1}^{M+1}(\mathbf{u}_j^\top \mathbf{y})\mathbf{u}_j.$$

In particular, the variance of the projected data is

$$\begin{aligned}
\frac{1}{N}\sum_{n=1}^{N}\left\|\sum_{j=1}^{M+1}(\mathbf{u}_j^\top \mathbf{x}_n)\mathbf{u}_j - \sum_{j=1}^{M+1}(\mathbf{u}_j^\top \overline{\mathbf{x}})\mathbf{u}_j\right\|_2^2 &= \frac{1}{N}\sum_{n=1}^{N}\left\|\Big[\sum_{j=1}^{M+1}\mathbf{u}_j\mathbf{u}_j^\top\Big](\mathbf{x}_n - \overline{\mathbf{x}})\right\|_2^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{M+1}\left\{\mathbf{u}_j^\top(\mathbf{x}_n - \overline{\mathbf{x}})\right\}^2 \\
&= \sum_{j=1}^{M+1}\mathbf{u}_j^\top \mathbf{S}\mathbf{u}_j,
\end{aligned}$$

where as before we take

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^\top$$

to be the data covariance matrix. Note that, as $\mathbf{u}_1, \ldots, \mathbf{u}_M$ are fixed, we are equivalently concerned with maximizing the single quantity

$$\mathbf{u}_{M+1}^\top \mathbf{S}\mathbf{u}_{M+1},$$

subject to the constraints $\mathbf{u}_{M+1}^\top \mathbf{u}_{M+1} = 1, \mathbf{u}_{M+1}^\top \mathbf{u}_j = 0$ for all $1 \leq j \leq M$. To solve the constrained optimization problem, we introduce the Lagrangian

$$\mathcal{L}(\mathbf{u}_{M+1}) = \mathbf{u}_{M+1}^\top \mathbf{S} \mathbf{u}_{M+1} - \sum_{j=1}^{M} \lambda_j \mathbf{u}_{M+1}^\top \mathbf{u}_j - \lambda_{M+1}\Big(\mathbf{u}_{M+1}^\top \mathbf{u}_{M+1} - 1\Big),$$

with undetermined coefficients λ. At any solution $\mathbf{u}_{M+1}^c$ to the optimization problem, there will be some $\lambda_1, \ldots, \lambda_{M+1}$ for which $\nabla \mathcal{L}(\mathbf{u}_{M+1}^c) = 0$. Computing,

$$\nabla \mathcal{L} = 2\mathbf{S}\mathbf{u}_{M+1} - \sum_{j=1}^{M} \lambda_j \mathbf{u}_j - 2\lambda_{M+1}\mathbf{u}_{M+1}.$$

In particular,

$$\mathbf{S}\mathbf{u}_{M+1}^c = \sum_{j=1}^{M} \frac{\lambda_j}{2} \mathbf{u}_j + \lambda_{M+1}\mathbf{u}_{M+1}^c.$$

Taking a transpose and multiplying on the right by some $\mathbf{u}_k$, $1 \leq k \leq M$,

$$(\mathbf{u}_{M+1}^c)^\top \mathbf{S}\mathbf{u}_k = \sum_{j=1}^{M} \frac{\lambda_j}{2} \mathbf{u}_j^\top \mathbf{u}_k + \lambda_{M+1}(\mathbf{u}_{M+1}^c)^\top \mathbf{u}_k.$$

We have used that $\mathbf{S} = \mathbf{S}^\top$, which is clear from its definition. By orthonormality,

$$(\mathbf{u}_{M+1}^c)^\top \mathbf{S}\mathbf{u}_k = \frac{\lambda_k}{2}.$$

On the other hand, our assumption is that $\mathbf{S}\mathbf{u}_k = \rho_k \mathbf{u}_k$ for some $\rho_k \in \mathbb{R}$, so

$$(\mathbf{u}_{M+1}^c)^\top \mathbf{S}\mathbf{u}_k = \rho_k (\mathbf{u}_{M+1}^c)^\top \mathbf{u}_k = 0.$$

Thus, $\lambda_k = 0$. Since $1 \leq k \leq M$ was arbitrary, we conclude that

$$\mathbf{S}\mathbf{u}_{M+1}^c = \lambda_{M+1}\mathbf{u}_{M+1}^c.$$

That is to say, $\mathbf{u}_{M+1}^c$ is a unit eigenvector of $\mathbf{S}$ orthogonal to all of $\mathbf{u}_1, \ldots, \mathbf{u}_M$.

Finally, if $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_D$ are the eigenvalues of $\mathbf{S}$, listed with multiplicity, then we have that $\mathbf{S}\mathbf{u}_j = \rho_j \mathbf{u}_j$ for all $1 \leq j \leq M$, and $\mathbf{S}\mathbf{u}_{M+1}^c = \rho_\iota \mathbf{u}_{M+1}^c$ for some $M+1 \leq \iota \leq D$. Recall that $\mathbf{u}_{M+1}^c$ is a solution to the optimization problem

$$\begin{aligned} &\underset{\mathbf{u}}{\operatorname{argmax}} \ \ \mathbf{u}^\top \mathbf{S}\mathbf{u} \\ &\text{subj. to} \ \ \|\mathbf{u}\| = 1,\ \mathbf{u}^\top \mathbf{u}_j = 0 \ \ \forall 1 \leq j \leq M. \end{aligned}$$

We have

$$(\mathbf{u}_{M+1}^c)^\top \mathbf{S}\mathbf{u}_{M+1}^c = \rho_\iota^2 \leq \rho_{M+1}^2.$$

Thus, the optimal value for $\mathbf{u}^\top \mathbf{S}\mathbf{u}$ is at most ρ_{M+1}^2. On the other hand, if we choose $\mathbf{u}$ to be an eigenvector of $\mathbf{S}$ corresponding to eigenvalue ρ_{M+1} which is orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_M$ and unit length, then we achieve the value ρ_{M+1}^2 while meeting the constraints. Thus, the optimization problem is solved by choosing $\mathbf{u}_{M+1}$ to be the vector described in the problem.

We have shown (by reference to the textbook) that the problem of projecting onto dimension 1 takes the desired form. We have shown also that, if the problem for dimension M is solved, then we can solve the problem for dimension $M+1$. Thus, by induction, we conclude the result for all dimensions $1 \leq M \leq D$. $\square$