Log-Sum-Exp

We have that $log(\sum_{i=0}^{k} exp(a_i - max_{j=0,k}\{a_j\})) + max_{j=0,k}\{a_j\} =$

$log(\sum_{i=0}^{k} exp(a_i - max_{j=0,k}\{a_j\})) + log(exp(max_{j=0,k}\{a_j\})) =$

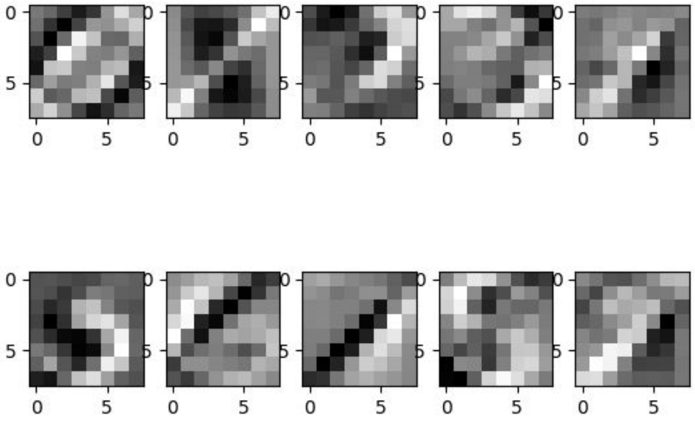
$log(\sum_{i=0}^{k} exp(a_i - max_{j=0,k}\{a_j\})exp(max_{j=0,k}\{a_j\})) =$ because sum of logs is log of product

$log(\sum_{i=0}^{k} exp(a_i - max_{j=0,k}\{a_j\} + (max_{j=0,k}\{a_j\})) =$

$log(\sum_{i=0}^{k} exp(a_i)$

Gaussian Discriminant Analysis

Average conditional log-likelihood for training set: -0.1246244366686302
Average conditional log-likelihood for test set: -0.1966732032552558
Training set accuracy: 0.9814285714285714
Test set accuracy: 0.97275



Dirichlet Distribution

a) Because $p(D|\theta) = \prod_{i=1}^{N} p(x^{(i)}|\theta)$, we have that

$p(\theta|D) \propto p(\theta)p(D|\theta) \propto (\theta_1^{\alpha_1 - 1}...\theta_K^{\alpha_k - 1})(\prod_{i=1}^{N}(\prod_{k=1}^{K} \theta_k^{x^{(i)}_k})) = (\theta_1^{\alpha_1 - 1}...\theta_K^{\alpha_k - 1})(\theta_1^{N_1}...\theta_K^{N_k})$, where N_i is the number of observations in the dataset that are of category i. This is then equivalent to $(\theta_1^{\alpha_1 - 1 + N_1}...\theta_K^{\alpha_k - 1 + N_k})$, which we recognize to be of the form of a Dirichlet distribution with parameters $\alpha_1 - 1 + N_1, ..., \alpha_k - 1 + N_k$.

b) From a), have that $log(p(\theta)p(D|\theta)) = (\alpha_1 - 1 + N_1)log(\theta_1) + ... + (\alpha_k - 1 + N_k)log(\theta_K) + c =$ $(\alpha_1 - 1 + N_1)log(1 - \sum_{j\neq 1}\theta_j) + \sum_{j\neq 1}(\alpha_j - 1 + N_j)log(\theta_j) + c$ for some constant c. Taking the derivative with respect to $\theta_i,\ i \neq 1$ and setting to 0, we get

$-(\alpha_1 - 1 + N_1)/(1 - \sum_{j\neq 1}\theta_j) + (\alpha_i - 1 + N_i)/\theta_i = 0 \rightarrow$ because the other terms in the latter summation don't contain the particular θ_i so they go to 0 when the derivative is taken

$-\theta_i(\alpha_1 - 1 + N_1) + (1 - \sum_{j\neq 1}\theta_j)(\alpha_i - 1 + N_i) = 0 \rightarrow$

$\theta_i(\alpha_1 - 1 + N_1) = (1 - \sum_{j\neq 1}\theta_j)(\alpha_i - 1 + N_i) \rightarrow$

$\theta_i/\theta_1 = (\alpha_i - 1 + N_i)/(\alpha_1 - 1 + N_1) \rightarrow$ because $\theta_1 = 1 - \sum_{j\neq 1}\theta_j$

$\sum_{i=1}^{K}\theta_i/\theta_1 = \sum_{i=1}^{K}(\alpha_i - 1 + N_i)/(\alpha_1 - 1 + N_1) \rightarrow$

$1/\theta_1 = \sum_{i=1}^{K}(\alpha_i - 1 + N_i)/(\alpha_1 - 1 + N_1) \rightarrow$

$\hat{\theta}_1 = (\alpha_1 - 1 + N_1)/\sum_{i=1}^{K}(\alpha_i - 1 + N_i)$

We can repeat this process for any category k, so we have that the kth component of out MAP estimate of $\vec{\theta}$ is $\hat{\theta}_k = (\alpha_k - 1 + N_k)/\sum_{i=1}^{K}(\alpha_i - 1 + N_i)$

c) From a), we have that $\theta|D$ ~Dirichlet$(\alpha_1 - 1 + N_1, ..., \alpha_k - 1 + N_k)$, so

$E(\theta_k|D) = \int_{-\infty}^{\infty}\theta_k p(\theta_k|D)d\theta_k = \alpha_k - 1 + N_k/\sum_{k'}(\alpha_{k'} - 1 + N_{k'}) \rightarrow$

$p(x_k{}^{(N+1)}|D) = \int_{-\infty}^{\infty}p(x_k{}^{(N+1)}|\theta_k)p(\theta_k|D)d\theta_k = \alpha_k - 1 + N_k/\sum_{k'}(\alpha_{k'} - 1 + N_{k'}) \rightarrow$

$p(x^{(N+1)}|D) = \int_{-\infty}^{\infty}p(x^{(N+1)}|\theta)p(\theta|D)d\theta = \prod_{k=1}^{K}(\alpha_k - 1 + N_k/\sum_{k'}(\alpha_{k'} - 1 + N_{k'}))$