CSC311 Homework 4

Eric Zhu

24/11/2020

Contents

Question 1:

Part A:

```
a = np.array([-5000000.0, -1010101010101, -650000])
```

```
b = np.array([10000000, 5000000, 6500000])
```

Part B:

Proof:

We want to prove that:

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

First we examine the left hand side of the equation above:

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

Applying the exponent and logarithm rules, we get:

$$\begin{aligned}
\log(\sum_{i=0}^{k} \exp(a_i - \max_{j=0}^{k} a_j)) + \max_{j=0}^{k} a_j &= \log(\sum_{i=0}^{k} \frac{\exp(a_i)}{\exp(\max_{j=0}^{k} a_j)} + \max_{j=0}^{k} a_j \\
&= \log(\sum_{i=0}^{k} \exp(a_i)) - \log(\exp(\max_{j=0}^{k} a_j)) + \max_{j=0}^{k} a_j \\
&= \log(\sum_{i=0}^{k} \exp(a_i)) - \max_{j=0}^{k} a_j + \max_{j=0}^{k} a_j \\
&= \log(\sum_{i=0}^{k} \exp(a_i))
\end{aligned}$$

Thus we have shown what we set out to prove.

■

Discussion of underflow/overflow:

Using the numerically stable version of `logsumexp`, we avoid overflow as the largest possible exponent is 0, and $\exp(0) = 1$ so we have at most $\log(k)$ (given k classes) for the log term, not `inf`. Conversely, we avoid underflow because supposing we have the largest possible difference of $a_i - \max_{j=0}^{k} a_j$, we find that $\exp(a_i - \max_{j=0}^{k} a_j))$ is possibly a really small positive number, and as such, the log of a small positive number (taking into considering floating point restrictions) is simply 0 but we add back on the max of a_i's, so we get an approximation of the `logsumexp`, not `-inf`.

Question 2:

Part A:

Average conditional log-likelihood (train): -0.12462443666862973

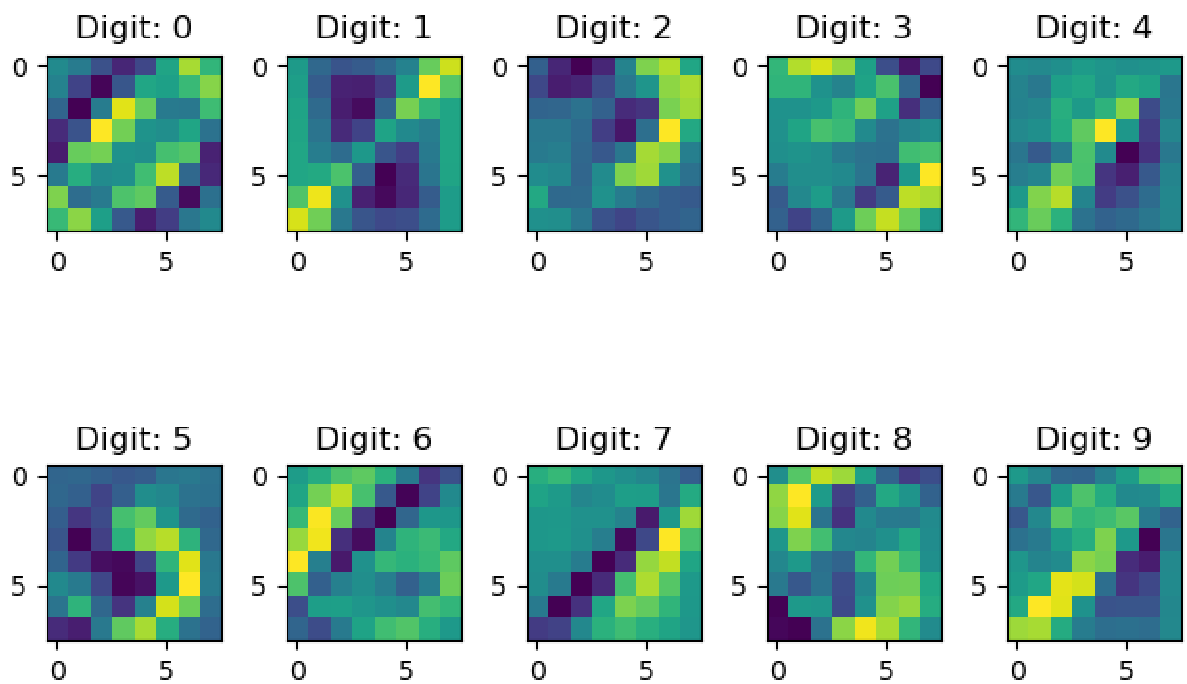
Average conditional log-likelihood (test): -0.19667320325525475

Part B:

Accuracy (train): 0.9814285714285714

Accuracy (test): 0.97275

Part C:



Question 3:

Part A:

We wish to derive the PDF of the posterior distribution, $p(\boldsymbol{\theta}|D)$.

Using Bayes rules we rewrite $p(\boldsymbol{\theta}|D)$ as $\frac{p(D|\boldsymbol{\theta})\cdot p(\boldsymbol{\theta})}{p(D)}$. Since $p(D)$ can be considered as a normalizing constant, we will calculate $p(D|\boldsymbol{\theta})\cdot p(\boldsymbol{\theta})$ instead. Thus we have:

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})\cdot p(\boldsymbol{\theta})$$

As we assume that the samples in D are i.i.d, we have that $p(D|\boldsymbol{\theta}) = p(x^{(1)}|\boldsymbol{\theta})p(x^{(2)}|\boldsymbol{\theta})...p(x^{(N)}|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{k=1}^{K}\theta_k^{x_k}$. Since a datapoint (sample) can only belong to 1 class, we have $\prod_{i=1}^{N}\theta_k^{x_k^{(i)}}$ will be exactly $\theta_k^{N_k}$, i.e., the value of θ for class k raised to the counts of samples belonging to class k in the dataset.To clarify, we are able to write this as $x_k = 1$ if and only if k is the correct class for x (otherwise $x_k = 0$) and θ_k is raised to the power of x_k. Thus, we can write $p(D|\boldsymbol{\theta})$ as:

$$p(D|\boldsymbol{\theta}) = \prod_{k=1}^{K}\theta_k^{N_k}$$

Next, substituting in the definition of $p(\boldsymbol{\theta})$ and $\prod_{k=1}^{K}\theta_k^{N_k}$ for $p(D|\boldsymbol{\theta})$, we have:

$$p(\boldsymbol{\theta}|D) \propto \prod_{k=1}^{K}\theta_k^{N_k}\cdot p(D|\boldsymbol{\theta})$$

It follows that:

$$\begin{aligned}
\prod_{k=1}^{K}\theta_k^{N_k}\cdot p(D|\boldsymbol{\theta}) &\propto \prod_{k=1}^{K}\theta_k^{N_k}\cdot(\theta_1^{a_1-1}\theta_2^{a_2-1}\theta_3^{a_3-1}...\theta_K^{a_K-1})\\
&= \prod_{k=1}^{K}\theta_k^{N_k}\prod_{k=1}^{K}\theta_k^{a_k-1}\\
&= \prod_{k=1}^{K}\theta_k^{N_k}\theta_k^{a_k-1}\\
&= \prod_{k=1}^{K}\theta_k^{N_k+a_k-1}
\end{aligned}$$

Thus, we have the PDF of the posterior distribution to be $\prod_{k=1}^{K}\theta_k^{N_k+a_k-1}$, which means $\boldsymbol{\theta}$ given the data D is Dirichlet distributed with parameters $(N_1+a_1, N_2+a_2, N_3+a_3, ..., N_K+a_K)$.

Part B:

We want to derive the MAP estimator for $\boldsymbol{\theta}$. Thus, we will start by considering the posterior distribution we derived in part A, i.e., $p(\boldsymbol{\theta}|D) = \prod_{k=1}^{K}\theta_k^{N_k+a_k-1}$. Then we define the log-likelihood function as $l(\boldsymbol{\theta}) = \log(p(\boldsymbol{\theta}|D)) = \sum_{k=1}^{K}(N_k+a_k-1)(\log(\theta_k))$.

To derive the MAP, we want to set $\partial l(\boldsymbol{\theta})/\partial\theta_k = 0$, but notice that the log-likelihood function only has one term after differentiation, i.e., $\frac{N_k+a_k-1}{\theta_k}$ because the other linear terms of the log-likelihood function such as $(N_1+a_1-1)\log(\theta_1)$ are constants.

Thus, we will use the Lagrangian method for constrained optimization. Here, our constraint is $\sum_k \theta_k = 1$. We will therefore optimize $l(\boldsymbol{\theta}) - \lambda(\sum_k \theta_k)$. Taking the partial derivative of that function with respect to θ_k (and set it to 0), we get:

$$\frac{N_k + a_k - 1}{\theta_k} - \lambda = 0 \iff \frac{N_k + a_k - 1}{\theta_k} = \lambda$$

We now need to solve for our Lagrange multiplier (lambda). We do so by substituting in $\frac{N_k+a_k-1}{\lambda}$ for θ_k into our constraint. We arrive at the equation:

$$\sum_k \frac{N_k + a_k - 1}{\lambda} = 1$$

Multiplying both sides by λ, we get:

$$\sum_k N_k + a_k - 1 = \lambda$$

Substituting in this value of lambda into our equation from above, we arrive at $\hat{\theta_k}$:

$$\hat{\theta_k} = \frac{N_k + a_k - 1}{\sum_k (N_k + a_k - 1)}$$

Part C:

We start with $p(\mathbf{x}^{(N+1)}|D) = \int p(\mathbf{x}^{(N+1)}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$. And we wish to find the probability of $\mathbf{x}^{(N+1)}$ being class k, i.e, $p(x_k^{(N+1)}|D) = \int p(x_k^{(N+1)}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$. From part A, we know that because $\mathbf{x}^{(N+1)}$ can only be one class (and $\mathbf{x}^{(N+1)}$ is only 1 sample), we have that $p(x_k|\boldsymbol{\theta}) = \theta_k$. Thus our integral becomes: $\int \theta_k p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$, and we realize that we have the definition of expectation, i.e., $\mathbb{E}[\theta_k|D]$. By the hint we have $\mathbb{E}[\theta_k|D] = \frac{N_k+a_k}{\sum_{k'}(N'_k+a'_k)}$ as $\boldsymbol{\theta} \sim \text{Dirichlet}(N_1 + a_1, N_2 + a_2, ..., N_k + a_k)$, which we derived in part A.