Homework 4 Solutions

1 Regression

1. *Regression with one predictor variable*

   (a) We will predict the mean of the y-values: $\widehat{y} = (1+3+4+6)/4 = 3.5$. The MSE of this prediction is exactly the variance of the y-values, namely:

   $$\text{MSE} = \frac{(1-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (6-3.5)^2}{4} = 3.25.$$

   (b) If we simply predict x, the MSE is

   $$\frac{1}{4}\sum_{i=1}^{4}(y^{(i)} - x^{(i)})^2 = \frac{1}{4}\left((1-1)^2 + (1-3)^2 + (4-4)^2 + (4-6)^2\right) = 2.$$

   (c) We saw in class that the MSE is minimized by choosing

   $$a = \frac{\sum_i (y^{(i)} - \overline{y})(x^{(i)} - \overline{x})}{\sum_i (x^{(i)} - \overline{x})^2}$$
   $$b = \overline{y} - a\overline{x}$$

   where $\overline{x}$ and $\overline{y}$ are the mean values of x and y, respectively. This works out to $a = 1, b = 1$; and thus the prediction on x is simply $x + 1$. The MSE of this predictor is:

   $$\frac{1}{4}\left(1^2 + 1^2 + 1^2 + 1^2\right) = 1.$$

2. *Lines through the origin*

   (a) The loss function is

   $$L(a) = \sum_{i=1}^{n}(y^{(i)} - ax^{(i)})^2$$

   (b) The derivative of this function is:

   $$\frac{dL}{da} = -2\sum_{i=1}^{n}(y^{(i)} - ax^{(i)})x^{(i)}.$$

   Setting this to zero yields

   $$a = \frac{\sum_{i=1}^{n} x^{(i)}y^{(i)}}{\sum_{i=1}^{n} {x^{(i)}}^2}.$$

3. The loss induced by a linear predictor $w \cdot x + b$ is

   $$L(w, b) = \sum_{i=1}^{n} |y^{(i)} - (w \cdot x^{(i)} + b)|.$$

4. Define

$$X = \begin{bmatrix} \leftarrow x^{(1)} \rightarrow \\ \leftarrow x^{(2)} \rightarrow \\ \vdots \\ \leftarrow x^{(n)} \rightarrow \end{bmatrix}$$

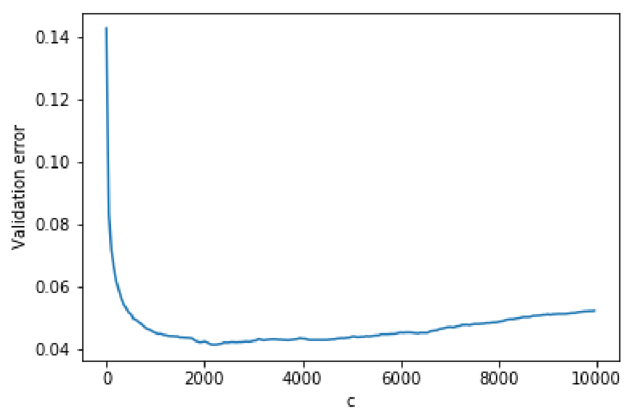
$$XX^T = \begin{bmatrix} x^{(1)} \cdot x^{(1)} & x^{(1)} \cdot x^{(2)} & \dots & x^{(1)} \cdot x^{(n)} \\ x^{(2)} \cdot x^{(1)} & x^{(2)} \cdot x^{(2)} & \dots & x^{(2)} \cdot x^{(n)} \\ x^{(n)} \cdot x^{(1)} & x^{(n)} \cdot x^{(2)} & \dots & x^{(n)} \cdot x^{(n)} \end{bmatrix}$$

2 Generative modeling

Pseudocode for training procedure:

- Load in the original training data matrix X and label vector y.
- Randomly split into validation set X_{valid}, y_{valid} of size 10,000 and training set X_{train}, y_{train}.
- For each digit $i = 0, 1, 2, \ldots$:
  - Calculate fraction of data points in training set with label i: π_i
  - Calculate mean of data points in training set with label i: $\mu^{(i)}$
  - Calculate covariance of data points in training set with label i: $\Sigma^{(i)}$
- For $c \in \{1, 51, 101, \ldots, 10001\}$:
  - Compute Gaussians $P_0 = \mathcal{N}(\mu^{(0)}, \Sigma^{(0)} + cI), \ldots, P_9 = \mathcal{N}(\mu^{(9)}, \Sigma^{(9)} + cI)$
  - Classify each validation point $x \in X_{\text{valid}}$ as the digit j which maximizes $\log \pi_j + \log P_j(x)$.
  - Compute the validation error (i.e. the fraction of validation points we misclassified).
- Select c^* to be the c which gave us the smallest validation error.

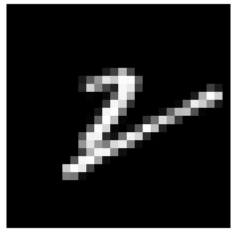
For a particular run, the above training procedure gives a validation error curve that looks like the following.



The c which achieves the minimum above is $c = 2151$. Note that your procedure may produce a different c due to the randomness in the choice of the validation/training split. The test error with this value is 0.0425.

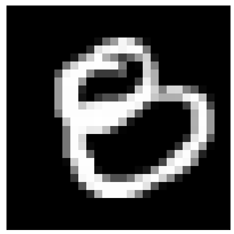
Now let's look at some randomly misclassified instances.

(a) The true label is 2, but it is predicted as 8.



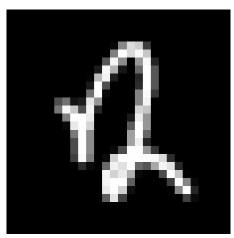
Pr(0|x)	Pr(1|x)	Pr(2|x)	Pr(3|x)	Pr(4|x)	Pr(5|x)	Pr(6|x)	Pr(7|x)	Pr(8|x)	Pr(9|x)
2.917e-19	7.710e-35	3.353e-13	5.478e-28	4.731e-25	1.851e-36	4.832e-17	7.189e-35	0.999	2.995e-34

(b) The true label is 8, but it is predicted as 0.



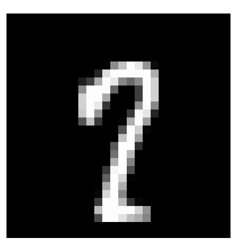
Pr(0|x)	Pr(1|x)	Pr(2|x)	Pr(3|x)	Pr(4|x)	Pr(5|x)	Pr(6|x)	Pr(7|x)	Pr(8|x)	Pr(9|x)
0.999	2.865e-276	6.603e-44	4.556e-7	1.093e-140	4.074e-45	4.951e-16	6.812e-161	4.820e-32	2.71e-139

(c) The true label is 2, but it is predicted as 4.



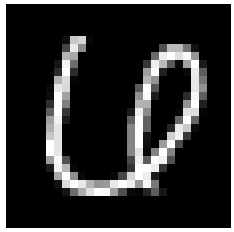
Pr(0|x)	Pr(1|x)	Pr(2|x)	Pr(3|x)	Pr(4|x)	Pr(5|x)	Pr(6|x)	Pr(7|x)	Pr(8|x)	Pr(9|x)
6.982e-35	4.744e-60	1.561e-12	2.882e-30	0.999	1.317e-27	8.505e-21	1.457e-36	2.046e-11	8.820e-23

(d) The true label is 7, but it is predicted as 9.



Pr(0|x)	Pr(1|x)	Pr(2|x)	Pr(3|x)	Pr(4|x)	Pr(5|x)	Pr(6|x)	Pr(7|x)	Pr(8|x)	Pr(9|x)
3.971e-98	6.371e-45	5.722e-65	8.567e-44	4.629e-18	1.671e-52	2.375e-107	0.157	3.690e-18	0.842

(e) The true label is 6, but it is predicted as 0.



$\Pr(0\|x)$	$\Pr(1\|x)$	$\Pr(2\|x)$	$\Pr(3\|x)$	$\Pr(4\|x)$	$\Pr(5\|x)$	$\Pr(6\|x)$	$\Pr(7\|x)$	$\Pr(8\|x)$	$\Pr(9\|x)$
0.999	5.719e-196	5.547e-41	2.758e-67	5.297e-70	5.333e-48	4.119e-37	2.928e-88	8.732e-52	3.977e-106