

Machine Learning Homework 1

Yussif Mustapha Tidoo

February 7, 2019

Question 1

Batch gradient descent computes the gradient of an example using the entire data set. i.e. each step of gradient descent uses all the examples in the training set to update the weights. This approach is computationally efficient. However, it requires loading the all the data set into the memory to make it available for the algorithm.

Stochastic gradient descent in contrast, updates the parameters of the model for each of the training example one-by-one. Since stochastic gradient descent updates the weights more frequently, it usually reaches convergence much faster than batch gradient descent. This approach is computa-

tionally expensive.

Mini-batch gradient descent is a combination of batch and stochastic gradient descent in that it splits the training data set into smaller batches and updates the parameters for each of the batches.

Question 2

Let **B** be a random variable representing that can take a value to be red box(r), green box(g) or blue box(b).

And **F** be a random variable that can take a value to be an apple(a), an orange(o) or a lime(l).

The probability of selecting the red box is:

$$p(B = r) = 0.2$$

The probability of selecting the green box is:

$$p(B = g) = 0.5$$

The probability of selecting the red box is:

$$p(B = b) = 0.3$$

We can write out all nine conditional probabilities for the type of fruit, given the selected box as:

$$p(F = a|B = r) = \frac{3}{10}$$

$$p(F = o|B = r) = \frac{4}{10}$$

$$p(F = l|B = r) = \frac{3}{10}$$

$$p(F = a|B = g) = \frac{3}{10}$$

$$p(F = o|B = g) = \frac{3}{10}$$

$$p(F = l|B = g) = \frac{4}{10}$$

$$p(F = a|B = b) = \frac{1}{3}$$

$$p(F = o|B = b) = \frac{2}{3}$$

$$p(F = l|B = b) = \frac{0}{3}$$

If a box is selected at random and a piece of fruit is picked from the box, the probability of selecting an apple is

$$p(F = a) = (0.3 * 0.2) + (0.3 * 0.5) + (0.3 * 0.333)$$

$$= 0.06 + 0.15 + 0.1$$

$$= 0.31$$

If we observe that the selected fruit is an orange, the probability that it came from the green box is:

$$p(B = g|F = o) = \frac{p(F = o|B = g)p(B = g)}{p(F = o)}$$

$$p(F = o) = p(F = o|B = r)p(B = r)$$

$$+ p(F = o|B = g)p(B = g)$$

$$+ p(F = o|B = b)p(B = b)$$

But

$$p(F = o) = (0.4 * 0.2) + (0.3 * 0.5) + (0.667 * 0.3)$$

$$= 0.08 + 0.15 + 0.2$$

$$= 0.43$$

Hence

$$p(B = g|F = o) = \frac{0.3 * 0.5}{0.43}$$

$$= 0.3488$$

Question 3

The distribution of c is: $P(c; u) = u^c(1 - u)^{1-c}$

The probability of the data $D = \{c^{(1)} \dots c^{(m)}\}$ is calculated as:

$$p(c^1, c^2, c^3 \dots c^m | u) = u^{c^1} (1-u)^{1-c^1} \cdot u^{c^2} (1-u)^{1-c^2} \cdot u^{c^3} (1-u)^{1-c^3} \dots u^{c^m} (1-u)^{1-c^m}$$

$$p(c; u) = \left(\prod_{i=1}^m u^{c^i} (1-u)^{1-c^i} \right)$$
$$p(c; u) = \left(u^{\sum_{i=1}^m c^i} \cdot (1-u)^{\sum_{i=1}^m 1-c^i} \right)$$

(a) The likelihood function is:

$$p(c; u) = \left(u^{\sum_{i=1}^m c^i} \cdot (1-u)^{\sum_{i=1}^m 1-c^i} \right)$$

To find the parameter u using Maximum Likelihood estimation, we will find the value of u at where the derivative of the likelihood function is zero. Thus, we are finding u that maximizes the likelihood function. I.e $\mu_{MLE} = \operatorname{argmax}_u p(c; \mu)$. The likelihood probability function above is difficult to differentiate, so it always simplified by taking the log.

$$\ln(p(c; \mu)) = \left(\ln(\mu^{\sum_{i=1}^m c^i} \cdot (1-\mu)^{\sum_{i=1}^m 1-c^i}) \right)$$

$$\ln(p(c; \mu)) = \left(\ln(\mu^{\sum_{i=1}^m c^i}) + \ln((1 - \mu)^{\sum_{i=1}^m (1 - c^i)}) \right)$$

$$\ln(p(c; \mu)) = \left(\sum_{i=1}^m c^i \ln(\mu) + \sum_{i=1}^m (1 - c^i) \ln((1 - \mu)) \right)$$

But

$$\sum_{i=1}^m c^i = m \bar{c}^i$$

$$\sum_{i=1}^m (1 - c^i) = m (1 - \bar{c}^i)$$

$$\ln(p(c; \mu)) = \left(m \bar{c}^i \ln(\mu) + m (1 - \bar{c}^i) \ln((1 - \mu)) \right)$$

$$\frac{\partial \ln(p(c; \mu))}{\partial \mu} = \frac{\partial}{\partial \mu} \left(m \bar{c}^i \ln(\mu) + m (1 - \bar{c}^i) \ln((1 - \mu)) \right)$$

$$\frac{\partial \ln(p(c; \mu))}{\partial \mu} = \frac{m \bar{c}^i}{\mu} - \frac{m (1 - \bar{c}^i)}{1 - \mu}$$

$$\frac{\partial \ln(p(c; \mu))}{\partial \mu} = 0$$

$$\frac{m \bar{c}^i}{\mu} - \frac{m (1 - \bar{c}^i)}{1 - \mu} = 0$$

$$\frac{\partial \ln(p(c; \mu))}{\partial \mu} = 0$$

$$\frac{m \bar{c}^i}{\mu} = \frac{m (1 - \bar{c}^i)}{1 - \mu}$$

$$\bar{c}^i (1 - \mu) = \mu (1 - \bar{c}^i)$$

$$\bar{c}^i - \mu \bar{c}^i = \mu - \mu \bar{c}^i$$

$$\mu = \bar{c}^i \Rightarrow \mu = \frac{\sum_{i=1}^m (c^i)}{m}$$

But

$$\sum_{i=1}^m (c^i) = H$$

Hence

$$u = \frac{H}{m}$$

(b) The parameter u is $\frac{H}{m}$

Prior probability, $p(\mu; \alpha) = \frac{1}{Z} \mu^{\alpha-1} (1 - \mu)^{\alpha-1}$

To derive the analytical expression of parameter μ using Maximum A Posteriori (MAP) estimation, we will maximize the log posterior in the Baye's rule. I.e. $\mu_{map} = \operatorname{argmax}_u p(u; c)$

Baye's rule: $p(u; c) = \frac{p(c; \mu)p(\mu; \alpha)}{p(c)}$

$$\operatorname{argmax}_u p(u; c) = \operatorname{argmax}_u \left(\frac{p(c; \mu)p(\mu; \alpha)}{p(c)} \right)$$

since $p(c)$ does not depend on μ , we can get rid of it.

$$\operatorname{argmax}_u p(u; c) = \operatorname{argmax}_u p(c; \mu)p(\mu; \alpha)$$

$$\ln(p(\mu; c)) = \ln(p(c; \mu)p(\mu; \alpha))$$

$$\frac{\partial \ln(p(\mu; c))}{\partial \mu} = \frac{\partial}{\partial u} (\ln(p(c; u)) + \ln(p(u; \alpha)))$$

$$\frac{\partial \ln(p(\mu; c))}{\partial \mu} = \frac{\partial}{\partial u} \ln(p(c; u)) + \frac{\partial}{\partial u} \ln(p(u; a))$$

From the maximum likelihood estimation, $\frac{\partial}{\partial u} \ln(p(c; u)) = \frac{m\bar{c}^i}{\mu} - \frac{m(1-\bar{c}^i)}{1-\mu}$. We can find $\frac{\partial}{\partial u} \ln(p(u; a))$ and add them up.

$$\frac{\partial}{\partial u} \ln(p(u; a)) = \frac{\partial}{\partial \mu} \left(\ln\left(\frac{1}{Z} \mu^{\alpha-1} (1-\mu)^{\alpha-1}\right) \right)$$

$$\frac{\partial}{\partial u} \ln(p(u; a)) = \frac{\partial}{\partial \mu} \left((\alpha-1) \ln\left(\frac{1}{Z} \mu\right) + (\alpha-1) \ln((1-\mu)) \right)$$

$$\frac{\partial}{\partial u} \ln(p(u; a)) = \frac{(a-1)}{u} - \frac{(a-1)}{(1-\mu)}$$

$$\frac{m\bar{c}^i}{\mu} - \frac{(a-1)}{(1-\mu)} = \frac{m(1-\bar{c}^i)}{1-\mu} - \frac{(a-1)}{u}$$

Question 4