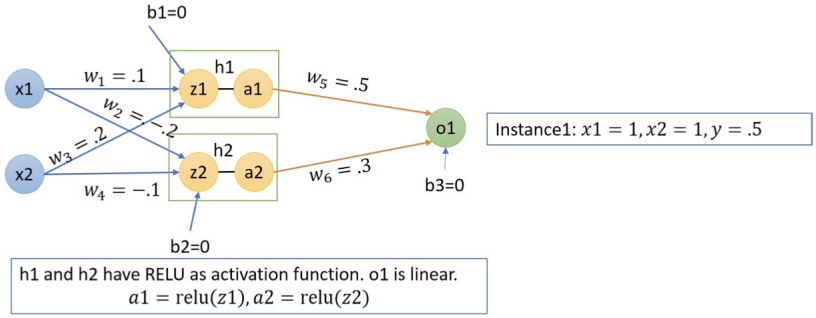
Tutorial ML

1 Artificial Neural Network

Consider the simple neural network drawn below for a regression task with two input units, one hidden layer with two neurons and RELU as the activation function, and a single output which is a linear unit.



Use the MSE loss function defined as $Loss(y, \hat{y}) = 1/2 \times (y - \hat{y})^2$.
The initial weight values associated with the network are- $w_1 = 0.1, w_2 = 0.2, w_3 = -0.2, w_4 = -0.1, w_5 = 0.5, w_6 = 0.3, b1 = b2 = b3 = 0$.

Recall, RELU(z)= max(0, z). Show how stochastic gradient descent using Instance1= $\{x_1 = 1, x_2 = 1, y = .5\}$ will update the network weights when the learning rate $\eta = 0.5$ by simulating a single pass of forward and backward propagation.

Solution: Instance1 : Input $x1 = 1, x2 = 1$; Output $y = 0.5$

$z1 = .3$ $\quad a1 = .3$
$z2 = -.3$ $\quad a2 = 0$
$o1 = .15$

$$L = \frac{1}{2} * (y - o1)^2$$

Loss, $L = .06125$

$$\frac{\partial L}{\partial o1} = -(y - o1) = -.35$$

$$\frac{\partial L}{\partial w_5} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial w_5} = -.35 * a1 = -.35 * .3 = -.105$$

$$\frac{\partial L}{\partial w_6} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial w_6} = -.35 * a2 = -.35 * 0 = 0$$

$$\frac{\partial L}{\partial b3} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial b3} = -.35 * 1 = -.35$$

Solution:

$$\frac{\partial L}{\partial a1} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial a1} = -.35 * w5 = -.175$$
$$\frac{\partial L}{\partial a2} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial a2} = -.35 * w6 = -.105$$

$$\frac{\partial L}{\partial z1} = \frac{\partial L}{\partial a1}\frac{\partial a1}{\partial z1} = -.175 * x1 = -.175$$
$$\frac{\partial L}{\partial z2} = \frac{\partial L}{\partial a2}\frac{\partial a2}{\partial z2} = 0$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_1} = -.175 * 1 = -.175$$
$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_3} = -.175 * x2 = -.175$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial w_2} = 0$$
$$\frac{\partial L}{\partial w_4} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial w_4} = 0$$

$$\frac{\partial L}{\partial b1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial b_1} = -.175 * 1 = -.175$$
$$\frac{\partial L}{\partial b2} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial b_2} = 0$$

$$w_1 = w_1 - \eta\frac{\partial L}{\partial w_1} = .1 + .5 * .175 = .1875$$
$$w_2 = w_2 - \eta\frac{\partial L}{\partial w_2} = -.2 + 0 = -.2$$
$$w_3 = w_3 - \eta\frac{\partial L}{\partial w_3} = .2 + 0.5 * .175 = .2875$$
$$w_4 = w_4 - \eta\frac{\partial L}{\partial w_4} = -.1 + 0 = -.1$$
$$w_5 = w_5 - \eta\frac{\partial L}{\partial w_5} = .5 + .5 * .105 = .5525$$
$$w_6 = w_6 - \eta\frac{\partial L}{\partial w_6} = .3 + 0 = .3$$
$$b1 = b1 - \eta\frac{\partial L}{\partial b1} = 0 + .5 * .175 = .0875$$
$$b2 = b2 - \eta\frac{\partial L}{\partial b2} = 0 + 0 = 0$$
$$b3 = b3 - \eta\frac{\partial L}{\partial b3} = 0 + .5 * .35 - .175$$

2 Expectation-Maximization Algorithm

Question
Consider two biased coins **A** and **B**. Let θ_A and θ_B be the probability of heads when Coin **A** and Coin **B** are tossed respectively.

Given below are 5 trials each containing 10 observations. Each trial is carried out by only one of the coins **A** or **B** but not both. Let these trials follow Binomial distribution. Let's take the number of heads as the event of success. *Note: Binomial distribution covers the case in which the exact sequence is unknown.*

Trial 1: H, T, T, T, H, H, T, H, T, H
Trial 2: H, H, H, H, T, H, H, H, H, H
Trial 3: H, T, H, H, H, H, H, T, H, H
Trial 4: H, T, H, T, T, T, H, H, T, T
Trial 5: T, H, H, H, T, H, H, H, T, H

Let $\theta_A = 0.6$ and $\theta_B = 0.5$ initially. Estimate the values of θ_A and θ_B using the Expectation Maximization algorithm for the aforementioned trials.

Also, Compare the values of θ_A and θ_B calculated using the EM algorithm with the case when we are given the information that Trial 1 and Trial 4 belong to Coin **A** and the remaining trials belong to Coin **B**.

Solution

The EM algorithm helps us estimate the parameters of a probability distribution.

Initial values for θ_A and θ_B are 0.6 and 0.5 respectively. It is given that trials follow the Binomial distribution i.e. $P(X) = \frac{n!}{(n-X)!X!} * p^X * q^{n-X}$ where n is the Total number of trials, X is the total number of successful events, p is the probability of a successful event and q is the probability of an unsuccessful event. But here the term $\frac{n!}{(n-X)!X!}$ is taken when we have no idea about how a sequence of events has taken place. But here, for a particular trial, we know the exact sequence of events and hence, we can drop this term.

Expectation step: Estimate the expected value for the hidden variable

$P(\frac{1^{st}trial}{1^{st}Coin}) = 1^{st}\,trail\,for\,1^{st}\,coin = (0.6)^5 * (0.4)^5 = 0.00079$
Similarly, $P(\frac{1^{st}trial}{2^{nd}Coin}) = 1^{st}\,trail\,for\,2^{nd}\,coin = (0.5)^5 * (0.5)^5 = 0.0009$
On normalizing these two probabilities, we approximately get:
$P(\frac{1^{st}Coin}{1^{st}Trial})$ = Probability(1^{st} coin used for 1^{st} trial = 0.45
$P(\frac{2^{nd}Coin}{1^{st}Trial})$ = Probability(2^{nd} coin used for 1^{st} trial = 0.55

Similarly, for the 2nd experiment, we have 9 Heads and 1 Tail. Hence:
Probability(1^{st} coin used for 2^{nd} experiment) = $(0.6)^9 * (0.4)^1 = 0.004$
Probability(2^{nd} coin used for 2^{nd} experiment) = $(0.5)^9 * (0.5)^1 = 0.0009$
On normalizing these two probabilities, we approximately get:
Probability(1^{st} coin used for 2^{nd} trial = 0.8
Probability(2^{nd} coin used for 2^{nd} trial = 0.2

Trial no.	**Coin A**	**Coin B**
1	0.45	0.55
2	0.8	0.2
3	0.73	0.27
4	0.35	0.65
5	0.65	0.35

Table 1: Normalized probabilities for Coin **A** and Coin **B** against every trial.

Now, we will multiply the Probability of the experiment belonging to the specific coin(calculated above) by the number of Heads and tails in the experiment.

Trial No.	**Coin A**	**Coin B**
1	0.45 * 5 H, 0.45 * 5 T	0.55 * 5 H, 0.55 * 5 T
2	0.8 * 9 H, 0.8 * 1 T	0.2 * 9 H, 0.2 * 1 T
3	0.73 * 8 H, 0.73 * 2 T	0.27 * 8 H, 0.27 * 2 T
4	0.35 * 4 H, 0.35 * 6 T	0.65 * 4 H, 0.65 * 6 T
5	0.65 * 7 H, 0.65 * 3 T	0.35 * 7 H, 0.35 * 3 T
Total	∼21.3H, 8.6 T	∼11.7 H, 8.4 H

Table 2: Result after multiplying the Probability of the experiment belonging to the specific coin by the number of Heads and Tails in the experiment.

Maximization Step: Optimize parameters using Maximum likelihood.

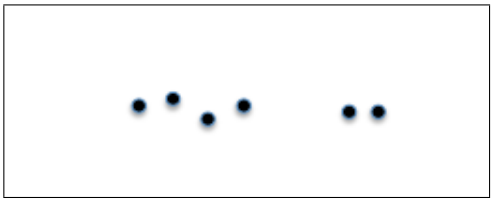
Now, update values for θ_A and θ_B are:
$\theta_A^1 = \frac{21.3}{21.3+8.6} = 0.71$ $\theta_B^1 = \frac{11.7}{11.7+8.4} = 0.58$

3 Gaussian Mixture Model

Question

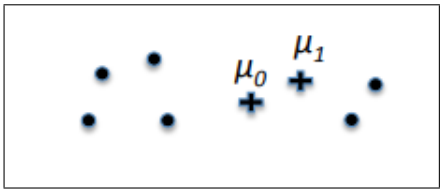
This question concerns training Gaussian Mixture Models (GMM's). Throughout, we will assume there are two Gaussian components in the GMM. We will use $\mu_0, \mu_1, \sigma_0, \sigma_1$ to define the means and variances of these two components, and will use π_0 and $(1 - \pi_1)$ to denote the mixture proportions of two Gaussians (i.e., $p(x) = \pi_0 N(\mu_0, \sigma_0 I) + (1 - \pi_0) N(\mu_0, \sigma_0 I)$). We will also use θ to refer to the entire collection of parameters $< \mu_0, \mu_1, \sigma_0, \sigma_1, \pi_0 >$ defining the mixture model $p(x)$.

3.1 Consider the set of training data below, and two clustering algorithms: K-Means and a Gaussian Mixture Model (GMM) trained using EM. Will these two clustering algorithms produce the same cluster centers (means) for this data set? If not then what will be position of centroids for the EM algorithm w.r.t. K-Means.



Answer: The difference lies in that k-means uses hard assignment of each point to a single cluster, whereas GMM uses soft assignment, where every point has a non-zero (though possibly small) probability of being in each cluster. So in k-means, the means of the clusters are determined by an average of the points assigned to that cluster, but in GMM the means of each cluster are (differently) weighted averages of all points. This has the effect of skewing the center of the left cluster to the right, and the center of the right cluster to the left. You could argue that this is a downside of the EM algorithm, that it still gives some weight to points that are clearly in the other cluster.

3.2 Consider applying EM to train a Gaussian Mixture Model (GMM) to cluster the data below into two clusters. The '+' points indicate the current means μ_0 and μ_1 of the two Gaussian mixture components after the kth iteration of EM.



3.2.a Draw on the figure the directions in which μ_0 and μ_1 will move during the next M-step.
Answer: μ_0 moves to the left, and μ_1 moves to the right.

3.2.b Will the marginal likelihood of the data, $\Pi_j P(x^j|\theta)$ increase or decrease on the next EM iteration? [note we use superscripts in this question to index training examples] Explain your reasoning in one sentence.
Answer: Increase. Each iteration of the EM algorithm increases to likelihood of the data unless you happen to be exactly at a local optimum.

3.2.c Will the estimate of π_0 increase or decrease on the next EM step? Explain your reasoning in one sentence.
Answer: Yes. π_0 is determined by adding the probabilities of all points that they are in cluster 1. In the current configuration, μ_1 is close enough to μ_0 that it will be stealing a lot of this probability mass so that π_0 and π_1 will be pretty close to each other.