ML Problems Part I

Ann He

July 2021

Intro

I'm on a journey... part of it is showing that ML math can be intuitive, personalized, and fun!

KL Divergence Chain Rule

We must show that

$$D_{KL}(P(X,Y) \parallel Q(X,Y)) = D_{KL}(P(X) \parallel Q(X)) + D_{KL}(P(Y|X) \parallel Q(Y|X)) \tag{1}$$

So...

$$\begin{aligned}
D_{KL}(P(X,Y)Q(X,Y)) &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)} \\
&= \sum_{x,y} P(x,y) \log P(x,y) - \sum_{x,y} P(x,y) \log Q(x,y) \\
&= \sum_{x,y} P(y|x)P(x) \log P(y|x) + \sum_{x,y} P(y|x)P(x) \log P(x) - \sum_{x,y} P(y|x)P(x) \log Q(y|x) - \sum_{x,y} P(y|x \\
&= \sum_{x} P(x)\big(\sum_{x,y} P(y|x) \log P(y|x)\big) - \sum_{x} P(x)\big(\sum_{x,y} P(y|x) \log Q(y|x)\big) + \sum_{x,y} P(y|x)P(x) \log P(x) \\
&= KL(P(Y|X) \parallel Q(Y|X)) + H(P(X), Q(X)) - H(P(X)) \\
&= KL(P(Y|X) \parallel Q(Y|X)) + D_{KL}(P(X) \parallel Q(X))
\end{aligned}$$

KL Divergence and MLE

We must show that

$$argmin_\theta D_{KL}(\hat{P} \parallel P_\theta) = argmax_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)}) \tag{2}$$

For me, it was important to distinguish iterating over x from the universe of all possible x's and iterating over over x from the dataset. Note that for x^i from the dataset,

$$P_\theta(x^i) = \sum_{x \in \mathcal{U}} P_\theta(x) 1\{x = x^i\} \tag{3}$$

The real weight of x^i is the sum of the weight of all x from the universe that equal it. (This is not useless pedantry!)

Ok, let's try it again

$$\begin{aligned}
min D_{KL}(\hat{P} \parallel P_\theta) &= \sum_{x \in \mathcal{U}} \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)} \\
&= -\sum_{x \in \mathcal{U}} (\frac{1}{n} \sum_{i=1}^{n} 1\{x = x^i\}) \log \frac{P_\theta(x)}{(\frac{1}{n} \sum_{i=1}^{n} 1\{x = x^i\})} \\
&= -\sum_{x \in \mathcal{U}} (\frac{1}{n} \sum_{i=1}^{n} 1\{x = x^i\}) \log P_\theta(x) + \sum_{x \in \mathcal{U}} (\frac{1}{n} \sum_{i=1}^{n} 1\{x = x^i\}) \log (\frac{1}{n} \sum_{i=1}^{n} 1\{x = x^i\}) \\
&= -\frac{1}{n} \sum_{x \in \mathcal{U}} \sum_{i=1}^{n} 1\{x = x^i\} \log P_\theta(x) \\
&= -\frac{1}{n} \sum_{x^i \in \mathcal{D}} \log P_\theta(x^i)
\end{aligned}$$

Which is equivalent to maximizing the log likelihood of the data.

EM for MAP Estimation

In this problem, we will (a) derive the EM update for MAP estimation and (b) prove that it monotonically increases the log(MAP).

Recall that we first derived EM for maximum likelihood estimation. The difference in this setting is that we have a prior on possible θ. Let's first restrict our attention to a single example x.

$$\begin{aligned}
\log \big(p(x|\theta)p(\theta)\big) &= \log \big[\big(\sum_{z \in z_x} p(x, z|\theta)\big)p(\theta)\big] \\
&= \log \big[\big(\sum_{z \in z_x} Q_x(z) \frac{p(x, z|\theta)}{Q_x(z)}\big)p(\theta)\big] \\
&= \log p(x, \theta) + \log \big(\sum_{z \in z_x} Q_x(z) \frac{p(x, z|\theta)}{Q_x(z)}\big) \\
&\geq \log p(x, \theta) + \sum_{z \in z_x} Q_x(z) \log \frac{p(x, z|\theta)}{Q_x(z)}
\end{aligned}$$

By Jensen's inequality. We now require $\frac{p(x,z|\theta)}{Q_x(z)} = c$, for equality. Here's one way to do it...

$$\lambda = \frac{p(x,z|\theta)}{Q_x(z)}$$
$$Q_x(z) = \frac{1}{\lambda} p(x,z|\theta)$$

We can set $\lambda = p(x \mid \theta) = \sum_z p(x,z|\theta)$, which is a constant since it depends on x, known, and not any z.

Now, let's argue monotonic improvement, which is a standard argument we already used before generically for EM for MLE.

We must show that $l(\theta^t) \leq l(\theta^{t+1})$ where l is the log(MAP).

$$\begin{aligned} l(\theta^{t+1}) &\geq ELBO(\theta^{t+1}) \\ &\geq ELBO(\theta^t) \\ &= l(\theta^t) \end{aligned}$$

Where the inequality between ELBOs comes from the fact that $\theta^{t+1} \coloneqq argmax_\theta ELBO(\theta^t)$ and the last inequality comes from the constant expectation in Jensen's Inequality.

EM Convergence

The EM algorithm monotonically optimizes the log likelihood of the data, i.e. we have $l(\theta^{t+1}) \geq l(\theta^t)$

Show that at θ^*, the point at which the EM algorithm reaches equality, or no longer makes improvement, we have reached a local maxima.

We must show that

$$\nabla_\theta l(\theta^*) = 0 \tag{4}$$

Whenever we get a problem like this, we can derive the statement (of equal to 0) on a single example x^i, and then conclude that the sum of 0-valued variables is 0.

$$\begin{aligned} l(x^i,\theta) &= \log p(x^i,\theta) \\ &= \log \sum_{z^i} p(z^i,x^i,\theta) \\ &= \log \sum_{z^i} Q_i(z^i) \frac{1}{Q_i(z^i)} p(z^i,x^i,\theta) \end{aligned}$$

Now...

$$\nabla_{\theta=\theta^*} l(x,\theta) = \frac{1}{p(x,z,\theta^*)} \sum_{z^i} Q_i(z^i) \frac{1}{Q_i(z^i)} \nabla_{\theta=\theta^*} p(x,z,\theta) = 0$$

We know this derivative of p is 0 by the fact that when EM reaches θ^*, the derivative of ELBO is 0. And if we use Q such that the ELBO bound is tight, then l by equality is also 0. Since $\frac{1}{p(x,z,\theta^*)}$ is a constant, then by implication, ∇l is 0.

PCA optimization, dual views

Reference: http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf

We must show that minimizing the average squared error between examples and their projections onto some unit vector u equals finding the first principal component of the data matrix.

Roughly, one can think of the conceptual path we take through this problem as MSE $\rightarrow$ mean/variance decomposition $\rightarrow$ covariance matrix $\rightarrow$ argue that you get the top eigenvector.

As usual, we restrict our analysis to a single example x, which makes the math easier, and then we generalize.

$$\begin{aligned}\left\|x-(w^Tx)w\right\|_2^2 &= x^Tx-2x^T(w^Tx)w+(w^Tx)^2w^Tw\\ &= \|x\|_2^2-(x^Tw)^2\end{aligned}$$

Since we are "minning" over w, we can ignore the first term. Let's add back all of the examples...

$$\begin{aligned}min-\sum_{i=1}^{n}(x_i^Tw)^2 &= min-\sum_{i=1}^{n}(w^Tx_i)(x_i^Tw)\\ &= max_{w:w^Tw=1}w^TXX^Tw\end{aligned}$$

We can see that w as the eigenvector with the largest eigenvalue maximizes this quantity.