Random Musings on a Sunday Morning
(extraordinarily basic stuff that’s worth reflecting upon now and then)

David Meyer
dmm@brocade.com

August 10, 2015

1 Introduction

One of the interesting things about machine learning is that it affords many different ways to look at the same data and hence underlying phenomena (aside: why exactly should this be?). For example, the Logistic Function is really a special case of a Conditional Random Field and PCA is a special case of an AutoEncoder. This document looks at a different correspondence. Here we look at why minimizing error (a sum), maximizing probability (a product), and minimizing energy (in an energy based model, for example, a Restricted Boltzmann Machine) are all really the same thing. Note that this document is likely to have many errors.

2 Minimizing Cost is a Sum

2.1 Linear Regression Cost Function

For linear regression, our hypothesis $h_\theta(x) = g(\theta^T x) = \theta_0 + \sum_{i=1}^n \theta_i x_i = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$. Here $g(z) = z$ (i.e., linear, but $g$ can be anything in other models). Now, given this hypothesis, our cost function can be written as a sum:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

The goal of machine learning then is to find the parameters $\theta$ such that cost/error function $J(\theta)$ is minimized (note that which error is minimized, and when, is a topic unto itself). For linear regression, Equation 1 is a convex optimization objective.
2.2 Logistic Regression Cost Function

For logistic regression, our hypothesis \( h_\theta(x) \) is slightly different, as shown in Equations 2, 3 and 4.

\[
\begin{align*}
    h_\theta(x) &= g(\theta^T x) \\
    g(z) &= \frac{1}{1 + e^{-z}}
\end{align*}
\]

Putting Equations 2 and 3 together we get

\[
    h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}
\]

While it seems like we could perhaps use the same cost function as we did in linear regression, it turns out that the linear regression cost function is non-convex when applied to logistic regression. As a result we typically use some version of cross-entropy as the cost function:

\[
    J(\theta) = \frac{1}{m} \left[ \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \right]
\]

In both cases we fit the parameters \( \theta \) to the model by minimizing a sum such as Equation 1 or 5.

3 Maximizing Probability is a Product

Basic assumption: all of this analysis depends on the assumption, as you will see, that the error \( \epsilon^{(i)} = y^{(i)} - \theta^T x^{(i)} \) is Gaussian, so Danger Will Robinson\(^2\). That said, assume that the target variables and the inputs are related via the equation

\[
    y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}
\]

where \( \epsilon^{(i)} \) is an error term that captures either unmodeled effects (such as if there are some important features that we left out of the model, or random noise). Interestingly this is intuitive as \( \epsilon^{(i)} = y^{(i)} - \theta^T x^{(i)} \). Assume also that the \( \epsilon^{(i)} \) are IID (Independent and

\(^{1}\)BTW, the logistic function is a special case of a Conditional Random Field

\(^{2}\)In reality it could be any parametric distribution. In this case I know what the simplifying assumptions are/mean so I can write the density down.
Identically Distributed) according to some Gaussian distribution with mean \( \mu = 0 \) and variance \( \sigma^2 \), i.e., \( N(0, \sigma^2) \). The general form of the probability density of \( N(\mu, \sigma^2) \) is:

\[
P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x - \mu)^2}{2\sigma^2}}
\] (7)

Given these assumptions, we can write down the density of \( \epsilon^{(i)} \) as follows:

\[
p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(\epsilon^{(i)})^2}{2\sigma^2}}
\] (8)

so that

\[
p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}
\] (9)

Rewriting this in vector/sum notation

\[
p(y \mid X, \theta, \sigma) = (2\pi\sigma^2)^{-m/2} e^{\frac{-1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2}
\] (10)

and taking advantage of the fact that the sum of exponential powers is product of exponentials\(^2\) we get

\[
p(y \mid X, \theta, \sigma) = \prod_{i=1}^{m} e^{\frac{-1}{2\sigma^2} (y^{(i)} - \theta^T x^{(i)})^2}
\] (11)

Note that the likelihood of a set of parameter values \( \theta \) given outcomes \( x \) is equal to the probability of those observed outcomes given those parameter values, that is

\[
\mathcal{L}(\theta \mid x) = P(x \mid \theta)
\] (12)

In the discrete case

\[
\mathcal{L}(\theta \mid x) = p_{\theta}(x) = P_{\theta}(X = x)
\] (13)

Thus we can see that Equation (11) is essentially the "likelihood of \( \theta \) given \( X \), and can be rewritten as

\[
\mathcal{L}(\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)
\] (14)

\(^3\) \( \exp(x) \) is defined to be \( e^x \)

\(^4\) \( a^{b+c} = a^b \times a^c \)
which familiarly equals

\[ \mathcal{L}(\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right) \] (15)

That is, maximizing the Likelyhood is a **product**.

### 4 Minimizing Energy

So we’ve seen that minimizing error (a sum) is roughly equivalent to maximizing the probability (or likelyhood), a product. It turns out that minimizing energy in a physical system is the same thing! In this section we’ll look at minimizing energy functions such as used by Restricted Boltzmann Machines (RBMs). Energy-based probabilistic models (e.g., RBMs) define a probability distribution through an energy function, as follows:

\[ p(x) = \frac{e^{-E(x)}}{Z} \] (16)

where the normalizing factor \( Z \), called a partition function by analogy to physical systems and is typically computationally intractable, is defined as follows

\[ Z = \sum_{i=1}^{n} e^{-E(x)} \] (17)
References