Random Musings on a Sunday Morning

(extraordinarily basic stuff that's worth reflecting upon now and then)

David Meyer

dmm@{1-4-5.net,uoregon.edu}

13 Oct 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 a Linear AutoEncoder (with no regularization). This document looks at a different correspondence. Here well 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* $h_\theta(\boldsymbol{x}) = g(\boldsymbol{\theta}^T\boldsymbol{x})$, where $g(z) = z$ and $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are $\mathrm{N} \times 1$ column vectors (hence $\boldsymbol{\theta}^T$ is a $1 \times \mathrm{N}$ row vector)[1]:

[1] While $g(z)$ is linear here, in other models g can be a non-linearity such as the *sigmoid* function.

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$\boldsymbol{\theta}^T = \begin{bmatrix} \theta_1, \theta_2, \cdots, \theta_n \end{bmatrix}$$

which implies that

$$\boldsymbol{\theta}^T \boldsymbol{x} = \begin{bmatrix} \theta_1, \theta_2, \cdots, \theta_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} \theta_i x_i = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{1}$$

Note that Equation 1 is the *algebraic interpretation* of the dot product of $\boldsymbol{\theta}^T$ and $\boldsymbol{x}$ (also called the inner or scalar product). An alternate, *geometric interpretation* of the dot product in Euclidian space, where a geometrical object possesses both a magnitude and a direction is

$$\boldsymbol{\theta}^T \boldsymbol{x} = \|\boldsymbol{\theta}^T\| \|\boldsymbol{x}\| \cos(\alpha) \tag{2}$$

where α is the is the angle between the vectors $\boldsymbol{\theta}^T$ and $\boldsymbol{x}$, and $\|\boldsymbol{\theta}^T\|$ and $\|\boldsymbol{x}\|$ are the magnitudes of the vectors $\boldsymbol{\theta}^T$ and $\boldsymbol{x}$ respectively (that is, the L^2-norm[2]). Note that Equation 2 implies that

$$\cos(\alpha) = \frac{\boldsymbol{\theta}^T \boldsymbol{x}}{\|\boldsymbol{\theta}^T\| \|\boldsymbol{x}\|} = \frac{\sum_{i=1}^{n} \theta_i^T x_i}{\sqrt{\sum_{i=1}^{n} (\theta_i^T)^2} \times \sqrt{\sum_{i=1}^{n} (x_i)^2}} \tag{3}$$

[2]The general form of the p-norm is $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$

which is also known as the *cosine similarity* between $\boldsymbol{\theta}^T$ and $\boldsymbol{x}$[1].

Getting back to our hypothesis, we see that $h_\theta(\boldsymbol{x}) = g(\boldsymbol{\theta}^T\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$. Note that by convention θ_0 is treated specially; in particular, it is not part of the vector $\boldsymbol{\theta}$. Now, given this hypothesis, our cost function can be written as a *sum*:

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{4}$$

The goal of machine learning then is to find the parameters $\boldsymbol{\theta}$ such that loss/error function $J(\boldsymbol{\theta})$ is minimized (note that which error is minimized, and when, is a topic unto itself). For linear regression, Equation 4 is a *convex* optimization objective.

2.2 Logistic Regression Cost Function

For logistic regression, our hypothesis $h_\theta(\boldsymbol{x})$ is slightly different, as shown in Equations 5, 6 and 7.

$$h_\theta(\boldsymbol{x}) = g(\boldsymbol{\theta}^T\boldsymbol{x}) \tag{5}$$

Here $g(z)$ is the *logistic* or *sigmoid* function[3], and is defined as follows

$$g(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

Putting Equations 5 and 6 together we get

$$h_\theta(\boldsymbol{x}) = g(\boldsymbol{\theta}^T\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T\boldsymbol{x}}} \tag{7}$$

While it seems like we could perhaps use the same loss function as we did in linear regression (Equation 4), it turns out that this loss function is *non-convex* when applied to logistic regression, i.e., when $g(z) = \frac{1}{1+e^{-z}}$. As as result we typically use some version of *cross-entropy* as the loss function:

$$J(\boldsymbol{\theta}) = \frac{1}{n}\left[\sum_{i=1}^{n} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right] \tag{8}$$

The regularized version of Equation 8 is

$$J(\boldsymbol{\theta}) = \frac{1}{n}\left[\sum_{i=1}^{n} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right] + \frac{\lambda}{2n}\sum_{j=1}^{n}\theta_j^2 \tag{9}$$

[3]BTW, the logistic function is a special case of a *Conditional Random Field*

2.3 Deriving the Loss/Error function for Logistic Regression

Recall that this loss function linear regression was

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2 \tag{10}$$

and that this loss function was a *non-convex* optimization objective for logistic regression. To find a convex loss function for logistic regression we first define a $Cost$ function:

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{11}$$

so that

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} Cost(h_\theta(x^{(i)}), y^{(i)}) \tag{12}$$

Now we can ask what our $Cost$ function look like. Well, if we predict 1 and $y = 1$, the the cost should be close to 0 (because we predicted the right value). Alternatively, if we predict 0 and $y = 1$, then the cost should converge on ∞ (that is, we penalize the prediction). On the other hand, if we predict 0 and $y = 0$, the the cost should again be close to close to 0 as we predicted the right value. Similarly, if we predict 1 and $y =$, the cost should again converge on ∞. These two cases are captured in Equation 13.

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases} \tag{13}$$

Note that if you combine the two cases in Equation 13 together, you get Equation 8, the *cross entropy*.

In any event, in both cases we fit the parameters $\boldsymbol{\theta}$ to the model by minimizing a **sum** such as Equation 4 or 8.

3 Maximizing Probability is a Product

Basic assumption: all of this analysis depends on the assumption, as you will see, that the error $\epsilon^{(i)}$ is Gaussian, so Danger Will Robinson![4] That said, assume that the target variables and the inputs are related via the equation

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)} \tag{14}$$

[4] 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.

and thus

$$e^{(i)} = y^{(i)} - \theta^T x^{(i)} \tag{15}$$

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 just random noise). Interestingly this is intuitive as $\epsilon^{(i)} = y^{(i)} - \theta^T x^{(i)}$. Assume also that the $\epsilon^{(i)}$'s are IID (Independent and Identically Distributed) according to some Gaussian distribution with mean $\mu = 0$ and variance σ^2, i.e., $\mathcal{N}(0, \sigma^2)$ (note that this argument also relies on the *Central Limit Theorem* and the *Law of Large Numbers* [2]). We typically use the notation $X \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that the random variable X is normally distributed with mean μ and variance σ^2 (note that if $X \sim \mathcal{N}(0, 1)$ we say that X follows the *standard normal* distribution).

The general form of the probability density function (pdf) of $\mathcal{N}(\mu, \sigma^2)$ is[5]

$$P(x) = f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right], \quad -\infty < x < \infty \tag{16}$$

Here $\mu = \mathbb{E}[X]$, i.e., the mean (and mode), $\sigma^2 = \text{var}[X]$, and $\sqrt{2\pi\sigma^2}$ is a normalization constant that ensures that the density f integrates to 1.

Now, assuming $e^{(i)} \sim \mathcal{N}(0, \sigma^2)$, we can write the density of $\epsilon^{(i)}$ as[6]:

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(\epsilon^{(i)})^2}{2\sigma^2} \right] \tag{17}$$

so that

$$p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right] \tag{18}$$

Rewriting this in vector/sum notation

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 \right] \tag{19}$$

[5] $exp(x)$ is defined to be e^x

[6] We can always "normalize" our data so that $\mu = 0$ by computing the z score, $z = \frac{x-\mu}{\sigma}$

and taking advantage of the fact that the sum of exponential powers is product of exponentials[7] we get

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} exp\left[-\frac{1}{2\sigma^2}(y^{(i)} - \theta^T x^{(i)})^2\right] \tag{20}$$

Note that the *likelyhood* of a set of parameter values θ given outcomes x is equal to the probability of those observed outcomes given those parameter values, that is

$$\mathcal{L}(\theta \mid x) = P(x \mid \theta) \tag{21}$$

In the discrete case

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x) \tag{22}$$

Thus we can see that Equation 20 is essentially the likelihood of θ given X, and can be rewritten as

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \tag{23}$$

which familiarly implies

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right] \tag{24}$$

That is, maximizing the likelihood is a **product**. Going the other way, maximizing the log likelihood $\ell(\theta)$ gives

$$\begin{aligned}
\ell(\theta) &= \log \mathcal{L}(\theta) && (25)\\
&= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right] && (26)\\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right] && (27)\\
&= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 && (28)
\end{aligned}$$

Hence maximizing $\ell(\theta)$ gives the same answer as minimizing

$$\frac{1}{2} \sum_{i=1}^{n} (y^{(i)} - \theta^T x^{(i)})^2 \tag{29}$$

[7] $a^{(b+c)} = a^b \times a^c$

which turns out to be the original least squares cost function $J(\boldsymbol{\theta})$ (Equation 4).

4 Minimizing Energy

So we've seen that minimizing error (a sum) is roughly equivalent to maximizing the probability (or likelihood), a product. It turns out that minimizing energy in a physical system such as a system of springs 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:

$$p(x) = \frac{e^{-E(x)}}{Z} \tag{30}$$

where the normalizing factor Z, called a partition function and is typically computationally intractable. Z is defined as follows

$$Z = \sum_{i=1} e^{-E(x)} \tag{31}$$

Using these definitions we can write the expression for the *likelihood* $\mathcal{L}$ of θ and $\mathcal{D}$ as

$$\mathcal{L}(\theta, \mathcal{D}) = \frac{1}{N} \sum_{x^{(i)} \in \mathcal{D}} \log p(x^{(i)}) \tag{32}$$

$$\ell(\theta, \mathcal{D}) = -\mathcal{L}(\theta, \mathcal{D}) \tag{33}$$

which we can minimize using stochastic gradient descent (SGD) where the gradient is $-\frac{\partial \log p(x^{(i)}}{\partial \theta}$ and $\boldsymbol{\theta}$ are the model parameters.

References

[1] The Wikipedia. Cosine similarity.

[2] The Wikipedia. Law of large numbers.