Notes on some basic probability stuff

David Meyer

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

September 30, 2014

1 Introduction

Note well: There are likely to be many mistakes in this document. That said...

Much of what is described here follows from two simple rules:

$$\text{Sum Rule:} \qquad P(\mathcal{X}) = \sum_{y} P(\mathcal{X}, \mathcal{Y}) \tag{1}$$

$$\text{Product Rule:} \qquad P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X}|\mathcal{Y})P(\mathcal{Y}) \tag{2}$$

- The Sum Rule is sometimes called marginalization
- The Product Rule is part of the proof of the Hammersley-Clifford Theorem

Remember also that $\mathcal{X} = \{x_i\}_{i=1}^{|\mathcal{X}|}$, where each x_i is a realization of the random variable x[1]. Lets also say that the set Θ of probability distribution parameters can be used to explain the *evidence* $\mathcal{X}$. Then we say that the "manner in which the evidence $\mathcal{X}$ depends on the parameters Θ" is the *observation model*. The analytic form of the observation model is the likelihood $P(\mathcal{X}|\Theta)$.

[1] Each observation x_i is, in general, a data point in a multidimensional space.

2 Estimating the parameters Θ with Bayes' Theorem

Note that

$$P(\Theta, \mathcal{X}) = P(\mathcal{X}, \Theta) \tag{3}$$
$$P(\Theta, \mathcal{X}) = P(\Theta|\mathcal{X})P(\mathcal{X}) \tag{4}$$
$$P(\mathcal{X}, \Theta) = P(\mathcal{X}|\Theta)P(\Theta) \tag{5}$$
$$P(\Theta|\mathcal{X})P(\mathcal{X}) = P(\mathcal{X}|\Theta)P(\Theta) \tag{6}$$

Solving for $P(\Theta|\mathcal{X})$ we get Bayes' Theorem

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)P(\Theta)}{P(\mathcal{X})} \tag{7}$$
$$\tag{8}$$

Said another way

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \tag{9}$$

You might also see Bayes' Theorem written using the *Law of Total Probability*[2] which is sometimes written as follows:

$$P(A) = \sum_n P(A \cap B_n) \qquad \text{\# by the } \textit{Sum Rule} \text{ (Equation 1)} \tag{10}$$
$$= \sum_n P(A, B_n) \qquad \text{\# in the notation used in Equation 1} \tag{11}$$
$$= \sum_n P(A|B_n)P(B_n) \qquad \text{\# by the } \textit{Product Rule} \text{ (Equation 2)} \tag{12}$$

so that the posterior distribution $P(\mathcal{C}_1|\mathbf{x})$ for two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ given input vector $\mathbf{x}$ would look like

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \tag{13}$$

[2]The Law of Total Probability is a combination of the Sum and Product Rules

Interestingly, the posterior distribution is related to logistic regression as follows: First recall that the posterior $P(\mathcal{C}_1|\mathbf{x})$ is

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \tag{14}$$

Now, if we set

$$a = \ln \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \tag{15}$$

we can see that

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1 + e^{-a}} = \sigma(a) \tag{16}$$

that is, the sigmoid function.

2.1 Maximum Likelihood Estimation (MLE)

Given all of that, for the MLE we seek the value of Θ that maximizes the likelihood $P(\mathcal{X}|\Theta)$ for our observations $\mathcal{X}$. Remembering that $\mathcal{X} = \{x_1, x_2, \ldots\}$ and that the x_i are iid, the value of Θ we seek maximizes

$$\prod_{x_i \in \mathcal{X}} P(x_i|\Theta) \tag{17}$$

Because of the product it is easier to use the log[3], we use the log likelihood $\mathcal{L}$:

$$\mathcal{L} = \sum_{x_i \in \mathcal{X}} \log P(x_i|\Theta) \tag{18}$$

and define $\hat{\Theta}_{ML}$ as follows:

$$\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{L} \tag{19}$$

[3]and since log(x) is monotonically increasing it doesn't effect the argmax

The maximization is obtained by (calculus tricks):

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \quad \forall \theta_i \in \Theta \tag{20}$$

Note finally that in a Generalized Linear Regression setting, we have

$$\eta = \mathbf{w}^T \mathbf{x} + b \tag{21}$$

$$p(y|\mathbf{x}) = p(y|g(\eta); \theta) \tag{22}$$

where $g(.)$ is an *inverse link function*, also referred to as an activation function. For example, if the link function is the logistic function, then the inverse link function $g(\eta) = \frac{1}{1+e^{-\eta}}$ and the negative log-likelihood $\mathcal{L}$ is

$$\mathcal{L} = -\log p(y|g(\eta); \theta) \tag{23}$$

2.2 Maximum a Posteriori (MAP) Estimation of Θ

Recall that

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)P(\Theta)}{P(\mathcal{X})} \tag{24}$$

We are seeking the value of Θ that maximizes $P(\Theta|\mathcal{X})$, so the solution can be stated as

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, P(\Theta|\mathcal{X}) \tag{25}$$

$$= \underset{\Theta}{\operatorname{argmax}}\, \frac{P(\mathcal{X}|\Theta) \cdot P(\Theta)}{P(\mathcal{X})} \tag{26}$$

However, since $P(\mathcal{X})$ does not depend on Θ, we can write

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, P(\mathcal{X}|\Theta) \cdot P(\Theta) \tag{27}$$

$$= \prod_{x_i \in \mathcal{X}} P(x_i|\Theta) \cdot P(\Theta) \tag{28}$$

If we again take the log, we get

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\operatorname{argmax}} \left(\sum_{x_i \in \mathcal{X}} \log P(x_i|\Theta) + \log P(\Theta) \right) \tag{29}$$

2.3 Notes

- Both MLE and MAP are point estimates for Θ (contrast probability distributions)
- MLE notoriously overfits
- MAP allows us to take into account knowledge about the prior (which is a sort of a regularizer)
- Bayesian estimation, by contrast, calculates the full posterior distribution $P(\Theta|\mathcal{X})$

2.4 Bayesian Estimation

Recall that Bayesian estimation calculates the full posterior distribution $P(\Theta|\mathcal{X})$, where

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\ P(\Theta)}{P(\mathcal{X})} \tag{30}$$

In this case, however, the denominator $P(\mathcal{X})$ cannot be ignored, and we know from the *sum* and *product* rules that

$$P(\mathcal{X}) = \int_{\Theta} P(\mathcal{X}, \Theta)\ d\Theta \tag{31}$$

$$= \int_{\Theta} P(\mathcal{X}|\Theta)\ P(\Theta)\ d\Theta \tag{32}$$

putting it all together we get

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\ P(\Theta)}{\int_{\Theta} P(\mathcal{X}|\Theta)\ P(\Theta)\ d\Theta} \tag{33}$$

If we want to be able to derive an algebraic form for the posterior $P(\Theta|\mathcal{X})$, the most challenging part will be finding the integral in the denominator. This is where the idea of *conjugate priors* and appoximate inference approaches (*Monte Carlo Integration* and *Variational Bayesian methods*[4]) are useful. Need to futher expand this...

[4]Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. They are typically used in complex statistical models consisting of observed variables (usually termed "data") as well as unknown parameters and latent variables, with various sorts of relationships among the three types of random variables, as might be described by a graphical model.

3 Monte Carlo Integration

Suppose we have a distribution $p(\theta)$ (perhaps a posterior) the we want to sample quantities of interest from. To do this analytically, we need to take an integral of the form

$$I = \int_{\Theta} g(\theta)\, p(\theta)\, d\theta \tag{34}$$

where $g(\theta)$ is some function of θ (typically $g(\theta) = \theta$ (the mean), etc). Need a deeper analysis here (note to self), but the punchline is that you can estimate I using *Monte Carlo Integration* as follows: Sample M values (θ^i) from $p(\theta)$ and calculate

$$\hat{I}_M = \frac{1}{M} \sum_{i=1}^{M} g(\theta^i) \tag{35}$$

Note that this works fine if the samples from $p(\theta)$ are iid[5] but if not, we can use a Markov Chain to draw "slightly dependent" samples and depend on the *Ergodic Theorem* (see Section 5.2).

4 Acknowledgements

[5] We know this by the Strong Law of Large Numbers, see Section 5.1.

References

5 Appendix

5.1 Strong Law of Large Numbers

Let $X_1, X_2, \ldots, X_M$ be a sequence of **independent** and **identically distributed** random variables, each having a finite mean $\mu_i = E[X_i]$.

Then with probability 1

$$\frac{1}{M}\sum_{i=1}^{M} X_i \to E[X] \tag{36}$$

as $M \to \infty$.

5.2 Ergodic Theorem

Let $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$ be M samples from a Markov chain that is *aperiodic*, *irreducible*, and *positive recurrent*[6], and $E[g(\theta)] < \infty$.

Then with probability 1

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta_i) \to E[g(\theta)] = \int_{\Theta} g(\theta)\, \pi(\theta)\, d\theta \tag{37}$$

as $M \to \infty$ and where π is the stationary distribution of the Markov chain.

[6] In this case, the chain is said to be *ergodic*.