STATISTICAL SCHOOLS OF THOUGHT HANDOUT (SLIDES 2)

COMP90051 STATISTICAL MACHINE LEARNING - SEMESTER 2, 2015

Abstract. This handout provides intuition into the topics covered in slides 2. The goal is to summarise intuition presented in lecture, but not written in slides, and perhaps not quite covered in the readings. In particular I've worked out in detail some of the "claims" presented in class.

1. Statistical Estimation

A central problem in statistics, and statistical machine learning, is to figure out the distribution that generates observed data. Carrying forward the loan application from class: let's write the vector of measurements representing the loan application (salary, number of dependents, etc.) as r.v. X and the eventual outcome of the loan (default or pay-off) as Y. If we knew the joint distribution $p(x, y)$ of these two r.v.'s then we could apply Bayes rule and *marginalisation* (summing out r.v.'s to get marginals - the distribution of individual r.v.'s) to determine other distributions in these variables of interest. For example, the conditional distribution $P(Y|X)$ would tell us: the probability of default/pay-off given the details of a loan application; with this we could decide (for example) to grant applications of higher likelihoods of profit for the bank.

In practice we don't always know what kind of distribution data $X_1, ..., X_n$ comes from, but we typically can (a) form an educated guess if we have an understanding of the underlying process generating the data - for example arrivals tend to follow Poisson or if data has a fixed number of values then we can use the multinomial distribution (b) we can evaluate our guess for the form of distribution using statistical goodness-of-fit tests - which we won't cover in this class or (c) we might start by using a distribution that is easier to work with like the normal, which is also common in nature. At the end of the day, we assume something about the family of distributions that data must have come from. In the *parametric approach*, we decide on a set of candidate distributions that differ only in a few parameters. We bundle the set of parameters (could be μ, σ for the normal or λ for the Poisson) as a parameter vector θ which must take its value from Θ. We also make the simplifying assumption that the data all comes from this same distribution, and is independent - that the data is *i.i.d.*

Our goal is to choose a $\hat{\theta} \in \Theta$ that fully specifies the distribution, based on the data. The chosen parameter (vector) value is called an *estimate* of the θ that generated the data.[1] The function of the data that generates our estimate, is called an *estimator*. The view we are taking here is that of the *frequentist* - there is some true (or "closest") parameter θ that we'd like to identify from data generated by p_θ (or nearly generated by this).

Example 1. Suppose we wanted to model the r.v. that the loan applicant is married - a yes/no choice and so a Bernoulli r.v. with some parameter $\theta \in [0, 1]$ the chance of "yes". Then given a sample of data of applicants being married or not $X_1, ..., X_n$ a decent estimator of the parameter might be $\frac{1}{n}\sum_{i=1}^{n} X_i$. Note that while the individual X_i's equal either 0 or 1 (single, married), the average could be any real in the interval $[0, 1]$.

2. Estimator Evaluation

A good practioner will not stop at defining an estimator for their problem. She will typically evaluate it on a hold-out test set (methdology for such empirical evaluation is covered in 90049 are mentioned in the optional review session). However for common estimators, we should like to have some provable justification that

Date: Last Updated: July 25, 2015.

[1] Our first assumption could have been wrong: maybe none of the p_θ's generated the data. In this case, think of the "true" θ as the "closest" possible fit to the underlying distribution.

they will work well, at least under "ideal" conditions: where the data is i.i.d. from a p_θ in our parametrised class of models.

2.1. **Bias and Variance.** Under the above ideal conditions, we'd like that on average our estimate $\hat{\theta}$ to be equal to the true θ. The degree that this is true is measured by *bias* as defined in class: $B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}(X_1, ..., X_n)] - \theta$. The bias might depend on the true θ that is generating the data, or it may be a constant. The first term measures the average value of the estimate when we repeatedly evaluate it on random ideal data. We then subtract from this the actual parameter to measure their distance. In class we calculated the bias of the sample average $\frac{1}{n}\sum_{i=1}^n X_i$ as zero for $N(\theta, 1)$ data. Note that the estimator is *unbiased* (has zero bias) no matter what θ was used to generate the data. Indeed the proof of unbiasedness goes through no matter what distribution is used, so long as the data is i.i.d.

Consider now two estimators $\hat{\theta}_1(.)$ and $\hat{\theta}_2(.)$ for the same parameter that are both unbiased. Which is better? A second important measure of estimator quality is its *variance*: this is just the usual definition of variance for a r.v.: $E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$ where I've dropped the depence of the estimate on the random data in the notation for readability. These two unbiased estimators may not have the same variance: the one with the smaller variance will be closer to being correct more often and is therefore superior.

2.2. **Asymptotic Measures.** As mentioned above, bias and variance may depend on the true θ generating the data. They may also depend on the data set size n: typically improving with growing n. While many estimators have bias or have less than minimal variance for a given size of dataset, we might then ask: are they eventually unbiased, or do they eventually have minimal variance? These two asymptotic properties of estimators are quantified by the notions of *consistency* and (asymptotic) *efficiency* mentioned in class.

3. Maximum-Likelihood Estimation

MLE is a general principle for estimation, that states

Estimate the $\hat{\theta}$ that maximises the likelihood of the observed data.

Likelihood of the observed data is measured by the *joint* probability of all the data together: $p_\theta(x_1, ..., x_n)$. But since we're assuming the ideal case, this potentially messy likelihood becomes the product of *marginals*: $\prod_{i=1}^n p_\theta(x_i)$. Thus in mathematical terms, we are to pick the parameter (vector) that maximises this product. Note that the mathematical expression we're optimising is in terms of data x_i's and parameter θ. But the maximisation is only in terms of the θ - the data is constant.

$$\hat{\theta}_{MLE}(X_1, ..., X_n) \quad := \quad \arg\max_{\theta \in \Theta} \prod_{i=1}^n p_\theta(X_i) \; . \tag{3.1}$$

3.1. **Log-space trick.** Optimising products is typically hard: recall from basic calculus that (in the case of unconstrained parameters that could be any real number) we proceed by taking the derivative of the function to be optimised with respect to the variable to select, and set this to zero. We then take the estimate as a root of this equation; we solve for the parameter. Derivatives of a large product as in Equation (3.1) can be quite messy. Instead we can take the log of the joint and maximise that: the log is a strictly increasing function, so maximising $f(x)$ produces the same x as maximising $\log(f(x))$.

$$\hat{\theta}_{MLE}(X_1, ..., X_n) \quad = \quad \arg\max_{\theta \in \Theta} \sum_{i=1}^n \log p_\theta(X_i)$$

What has happened here, is that the log of a product equals the sum of the log terms. Finally to maximise: the derivative of a sum, is the sum of derivatives (the derivative is a "linear operator"), which makes computing the max much, much easier.

Remark 2. This log-space trick is used in other context in statistical learning. For example, Markov models or hidden Markov models involve long products of small probabilities which can vanish when doing floating-point arithmetic. Instead it is common to work with the log-probabilities to.

Example 3. In class I mentioned that the maximum-likelihood estimator for the mean of a normal distribution with known variance (say 1) is the sample average. Let's go through this together now

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\mu\in\mathbf{R}} \sum_{i=1}^{n} \log Normal(X_i;\mu,1) \\ &= \arg\max_{\mu\in\mathbf{R}} \sum_{i=1}^{n} -(X_i-\mu)^2\end{aligned}$$

Note in the second line we dropped constants as they don't affect the optimising parameter value. To solve this (unconstrained) optimisation, we set the derivative (with respect to μ) to zero and solve for our parameter

$$\begin{aligned}0 &= \frac{\partial}{\partial\mu} \sum_{i=1}^{n} -(X_i-\hat{\mu})^2 \\ &= 2\sum_{i=1}^{n} (X_i-\hat{\mu})\end{aligned}$$

dividing both sides by 2 and rearranging, yields $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. (As a sanity check, we could also check that this optimum is indeed a maximum not a minimum - which indeed it is).

Exercise: what is the bias of this estimator?

4. Bayesian Statistics

The *Bayesian* is concerned with similar problems as the frequentist, but holds a different interpretation of probability - as a *degree of belief* - and as such considers unknown parameters not as objects holding some fixed value, but rather random variables themselves. There are statisticians and machine learning researchers who prescribe to one of these two philosophies and vehemently deny the validity of the other - frequentist vs Bayesian is oft jokingly described as a "religious" choice - but many researchers see merit in both frameworks and will pick one that provides effective modeling of the problem at hand.

Example 4. In class we had an example of a r.v. X which that is Normal with unknown mean μ and known variance 1. The pure Bayesian approach to inferring something about μ from an i.i.d. sample of X is not to calculate a fixed estimate (called a "point estimate") but rather to generate a distribution over μ representing the Bayesian's degree of belief for possible values of μ. Treating μ as a r.v., the Bayesian will begin her modeling exercise by placing a prior distribution $P(\mu)$ over the parameter before any data is collected at all. Let's use $Normal(0,1)$ for the prior. In this way, X is actually normal-distributed given μ so we say that the conditional $P(X|\mu)$ is $Normal(\mu,1)$; previously the frequentist was modeling X as unconditionally normal, with an unknown mean. Coming back to what the Bayesian does with a datum $X = x$: she updates the distribution on μ by calculating the posterior distribution. In class I gave the answer for when $x = 1$ let's see the working for general x:

$$
\begin{aligned}
P(\mu|X=x) &= \frac{P(X=x|\mu)P(\mu)}{P(X=x)} \\
&\propto P(X=x|\mu)P(\mu) \\
&\propto \exp\left(-\frac{(x-\mu)^2}{2}\right)\exp\left(-\frac{\mu^2}{2}\right) \\
&= \exp\left(-\frac{(x-\mu)^2+\mu^2}{2}\right) \\
&= \exp\left(-\frac{x^2-2x\mu+2\mu^2}{2}\right) \\
&= \exp\left(-\frac{0.5x^2-x\mu+\mu^2}{2\cdot 0.5}\right) \\
&= \exp\left(-\frac{(0.5x-\mu)^2+x^2/4}{2\cdot 0.5}\right) \\
&= \exp\left(-\frac{(0.5x-\mu)^2}{2\cdot 0.5}\right)\exp\left(-\frac{x^2}{4}\right) \\
&\propto \exp\left(-\frac{(0.5x-\mu)^2}{2\cdot 0.5}\right) \\
&\propto Normal(0.5x, 0.5)
\end{aligned}
$$

What we have done here is to apply Bayes rule, plugged in the forms of our conditional and prior distributions, then we collected terms, and recognised that we will end up with the form of a normal again. Once we have realised we are looking to write the result in this form, we can massage the expression so that we can read-off the parameters of this normal.

In detail, we first apply Bayes rule. We then drop the denominator which does not depend on μ. Next we plug in the densities for the conditional and prior again dropping constants: all that matters are terms that depend on μ as the rest are "normalisers" - once we know the final family of distribution for the posterior we can always look up the correct normalising constants. We next collect terms. At this point we are looking for what kind of distribution we are dealing with: this distribution is in μ - any x's are should be thought of as constants. It is clear that we have a distribution that is exponential in a quadratic in μ this must be a normal! We know that a $Normal(a, b^2)$ should have form $\exp\left(-\frac{(\mu-a)^2}{2b^2}\right)$ and so we perform several lines of algebra (essentially we are completing the square) to get to this form. Any left-over constants (terms not involving μ) from inside the exponential can come outside and then get dropped.

In this example we have updated the prior to the posterior. Notice that the mean for μ moved towards the datum we observed, and also notice that the variance for μ decreased as we are now a little more sure about μ's value, having gathered more evidence. This pattern would continue as we observe (and train on) more data.