CS 536 : Probability Notes - How does Data reflect Reality? 16:198:536

In general, we view 'data' as a collection of samples $X_1, X_2, \ldots, X_n$ drawn from some source. The individual X_i represent samples, or instances, or measurements, of some underlying 'ground truth' - for instance, the results of the flips of a coin reflecting the underlying probability of heads, or images of dogs as samples from the space of images of dogs. In practice, while we have access to the data generated from the underlying distribution, this ground truth itself is in many ways hidden from us - the only access we have to it is through the specific results or measurements $X_1, \ldots, X_n$. The language and mechanics of probability theory allows us to analyze this underlying distribution based on the concrete results we have.

Probability and Notation

It is convenient to define the following terms and notation: the *cumulative distribution function* or c.d.f. of a random measurement X is given by

$$F(x) = \mathbb{P}(X \leq x). \tag{1}$$

The function $F(x)$ quantifies how likely it is or how much belief we should assign to the result of a measurement being at most x. In general, it should be true that as $x \to \infty$, we should have that $F(x) \to 1$, and that as $x \to -\infty$, we should have that $F(x) \to 0$. The *probability density function* (p.d.f.) (or mass function, p.m.f. in the case of discrete valued random variables) essentially indicates how likely X is to fall in given ranges or on given values, in the following way:

$$\int_a^b f(x)dx = \mathbb{P}(a \leq X \leq b). \tag{2}$$

Note relatedly, that $F(b) - F(a) = \int_a^b f(x)dx$, and that $F'(x) = f(x)$.

The mean *expected value* of X or $\mathbb{E}[X]$ represents the 'belief-weighted average of X':

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx. \tag{3}$$

This is a useful measure of what values of X are likely or have reasonable belief to occur. But measurements that come out to exactly $\mathbb{E}[X]$ are rare - we measure the discrepancy or how much *variance* around the mean to expect via

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]. \tag{4}$$

Note an important relationship here: $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

In general, we denote the mean and variance as μ and σ^2 respectively.

Concentration Bounds

If X represents a measurement from some underlying probability distribution $\mathbb{P}$, then the mean and expected value μ, σ^2 represent theoretical properties of this distribution, rather than the specific measurement or data point X. A number of results connect a given measurement X to its theoretical properties, however.

In the case that $X > 0$, that is, X is always a strictly positive value, the **Markov Inequality** gives us

$$\mathbb{P}(X > t) \leq \frac{\mu}{t}. \tag{5}$$

Note, as $t \to \infty$, this gives us that $\mathbb{P}(X > t) \to 0$, but this is not surprising - the probability of getting arbitrarily large measurements should go to 0. But this inequality quantifies *how fast* this probability goes to 0, in that it is at least as fast as $O(1/t)$. In application, this tells us for instance that

$$\mathbb{P}(X > 2\mu) \leq \frac{1}{2}, \tag{6}$$

i.e., the likelihood of a measurement being greater than twice its expected value is at most $1/2$. That is, most of the time, we should not expect a measurement to deviate from its mean by more than a factor of 2.

Tighter results are available. For a general X, with no restriction on positivity, the **Chebychev Inequality** gives us that

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}. \tag{7}$$

This gives us a stronger result - that X should in fact concentrate at μ, and the likelihood of deviating away from there falls like $O(1/t^2)$. Hence, for example, the likelihood of a measurement deviating from the mean by two standard deviations is at most $1/4$:

$$\mathbb{P}(|X - \mu| > 2\sigma) \leq \frac{1}{4}. \tag{8}$$

So 75% of measurements for X should fall in the interval $[\mu - 2\sigma, \mu + 2\sigma]$.

It's worth noting that the Chebychev Inequality is a special case of the Markov Inequality, applied to bound the probability $\mathbb{P}\left((X - \mu)^2 > t^2\right)$.

Concentration of Averages

The previous results tell us how individual measurements relate to the underlying distribution - but suppose we wanted to *estimate* the value of μ. The typical estimate for the mean is of course the sample mean, i.e.,

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{9}$$

We then expect that we should have $\hat{\mu} \approx \mu$. But why is this valid? And can we be more precise? To begin with, it's worth noting the following. Based on usual relationships between mean and variance for i.i.d. random variables, we have

$$\begin{aligned} \mathbb{E}\left[\bar{X}_n\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu \\ \mathrm{Var}\left(\bar{X}_n\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned} \tag{10}$$

Applying Chebychev in this case, we see that (regardless of the underlying distribution!), we have

$$\mathbb{P}(|\hat{\mu} - \mu| > \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2} = \frac{1}{n}\left(\frac{\sigma}{\epsilon}\right)^2 = O\left(\frac{1}{n}\right). \tag{11}$$

Or in short, the likelihood of the estimator deviating from the 'true value' of μ by at least ϵ decreases like $O(1/n)$. As the number of samples we take goes up, there is an increasing likelihood that the value of the estimator is within ϵ of the true value of μ.

Tighter Concentration Bounds

In the case that we know more about the distribution of X, we can say more about the concentration of the average around the mean. Consider the case for instance where each $X_i \in \{0, 1\}$, for instance each X_i representing the success or failure of some test. Let $\mu = \mathbb{E}[X]$ in the usual way, essentially taking μ to represent the 'probability of success'.

In this case, we can state the following bound. For any $\lambda > 0$, we have:

$$\begin{aligned}
\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i > (\mu+\epsilon)n\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} \lambda X_i > \lambda(\mu+\epsilon)n\right) \\
&= \mathbb{P}\left(\exp\left(\sum_{i=1}^{n} \lambda X_i\right) > \exp\left(\lambda(\mu+\epsilon)n\right)\right) \\
&= \mathbb{P}\left(\prod_{i=1}^{n} e^{\lambda X_i} > e^{\lambda(\mu+\epsilon)n}\right) \\
&\leq \frac{\mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right]}{e^{\lambda(\mu+\epsilon)n}}.
\end{aligned} \tag{12}$$

The last step above is simply an application of the Markov Inequality. Observing that the X_i are independent however, that expected value actually factors into the product of expected values:

$$\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) \leq \frac{\prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right]}{e^{\lambda(\mu+\epsilon)n}}. \tag{13}$$

But given that all the measurements are i.i.d., all those expected values are the same:

$$\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]^n}{e^{\lambda(\mu+\epsilon)n}}. \tag{14}$$

For X distributed in this way, $\mathbb{E}\left[e^{\lambda X}\right] = e^0(1-\mu) + e^{\lambda * 1} * \mu = (1-\mu) + \mu e^{\lambda}$. Hence we get a final bound of

$$\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) \leq \left(\frac{(1-\mu) + \mu e^{\lambda}}{e^{\lambda(\mu+\epsilon)}}\right)^n. \tag{15}$$

Note: this bound is valid for all relevant $\mu, \epsilon, \lambda > 0$. But the bound is only *useful* if the quantity being raised to the n is less than 1, and thus the bound would shrink exponentially fast with n. By choosing μ to *maximize* the upper bound, we ensure that the bound will be valid for all possible μ we might encounter - a worst case upper bound. But we then still have the freedom to choose λ to try to minimize that worst case upper bound. In particular, for any λ, the bound is maximized taking

$$\mu^* = \frac{1}{\lambda} - \frac{1}{e^{\lambda} - 1}. \tag{16}$$

This yields a general bound of:

$$\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) \leq \left[\frac{1}{e} e^{\left(\frac{\lambda}{e^{\lambda}-1}\right)} \left(\frac{e^{\lambda}-1}{\lambda}\right) e^{-\lambda\epsilon}\right]^n. \tag{17}$$

This is a relatively nasty looking function - in moments like this, it is frequently useful to fall back to approximations, etc. It can be shown that

$$\frac{1}{e} e^{\left(\frac{\lambda}{e^{\lambda}-1}\right)} \left(\frac{e^{\lambda}-1}{\lambda}\right) \leq e^{\frac{1}{8}\lambda^2}, \tag{18}$$

hence

$$\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) \leq e^{\left(\frac{1}{8}\lambda^2 - \lambda\epsilon\right)n}. \tag{19}$$

Lastly, choosing λ to try to minimize this bound, the minimum occurs at $\lambda^* = 4\epsilon$, giving a final bound of

$$\mathbb{P}\left(\bar{X}_n > \mu + \epsilon\right) \leq e^{-2\epsilon^2 n}. \tag{20}$$

This is the so called 'Hoeffding Inequality'. It is straightforward to extend this bound to show as well that

$$\mathbb{P}\left(|\bar{X}_n - \mu| > \epsilon\right) \leq 2e^{-2\epsilon^2 n}. \tag{21}$$

Hence we get that the sample mean concentrates incredibly tightly around the true mean as a function of n.

One application of this kind of concentration result is the following: suppose you wanted to establish the true mean μ with high confidence and high accuracy, i.e., estimate the true mean μ within ϵ, with probability at least $1 - \delta$. In that case, you would need to take at least

$$n \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta} \tag{22}$$

many samples, so that

$$\mathbb{P}\left(|\bar{X}_n - \mu| > \epsilon\right) \leq 2e^{-2\epsilon^2\left(\frac{1}{2\epsilon^2} \ln \frac{2}{\delta}\right)} = 2e^{-\ln \frac{2}{\delta}} = \delta. \tag{23}$$

This illustrates a common result, that in order to establish a result with accuracy ϵ with confidence at least $1 - \delta$, we need at least

$$n = O\left(\frac{1}{\epsilon^2} \ln\left(\frac{1}{\delta}\right)\right) \tag{24}$$

many samples.