DS102 - Discussion 6

Wednesday, 9th October, 2019

In this section we prove several useful concentration inequalities. A concentration inequality is an inequality of the form $\mathbb{P}(X \geq t) \leq \delta$, or equivalently $\mathbb{P}(X < t) > 1 - \delta$, for some random variable X. The goal is to find the smallest δ that makes the inequality true for a fixed t (or the smallest t that makes the inequality true for a given δ).

There is a number of reasons why we care about concentration inequalities. For one, it is typically hard to design a procedure that will work correctly *always*. For example, it is impossible to guarantee that a self-driving car will recognize a stop sign with probability one. We *can* say, however, that it will recognize it with probability at least $1 - \delta$, for some tiny δ. Therefore, one motivation is being able to say $\mathbb{P}(\text{extreme events}) \leq \delta$, where an "extreme event" is $X \geq t$, for some relevant quantity X. Another motivation is in constructing confidence intervals. If a family of distributions is "well-concentrated", meaning that for a fixed t we can make δ very tiny, then we can give relatively small confidence intervals (meaning that we're more confident about the unknown parameter).

1. Prove Markov's inequality, which states that for all *non-negative* random variables X, $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$, for all $t > 0$.

Solution: For simplicity, we assume that X has a density $p(x)$. By definition of expectation, we have

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x p(x) dx \\ &= \int_0^t x p(x) dx + \int_t^\infty x p(x) dx\end{aligned}$$

Now we focus on the second term. Since it only considers $x \geq t$, we can write the following lower bound

$$\mathbb{E}[X] = \int_0^t x p(x) dx + \int_t^\infty x p(x) dx \geq \int_0^t x p(x) dx + t \int_t^\infty p(x) dx.$$

Since the first term is non-negative (and notice that this is due to X being a non-negative random variable), we can ignore it to get

$$\mathbb{E}[X] \geq t \int_t^\infty p(x) dx = t\mathbb{P}(X \geq t),$$

where is the last step we use the definition of the density function. Rearranging gives

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

2. Prove the Chernoff bound, which states that for any random variable X,

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \inf_{\lambda \geq 0} e^{-\lambda t - \lambda \mathbb{E}[X]} \mathbb{E}[e^{\lambda X}],$$

for all $t > 0$. The function $\psi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ is called the *moment-generating function* of X. The moment-generating funcition (MGF) is important because it uniquely determines a distribution (just like a CDF, or density/PMF).

Solution: We will use the following fact: for any two random variables Y, Z,

$$\mathbb{P}(Y \geq Z) \leq \mathbb{P}(g(Y) \geq g(Z)),$$

for any non-decreasing function g. Moreover, for a strictly increasing function g, the above inequality becomes an equality. Let λ be an arbitrary non-negative constant, and let $Y := X - \mathbb{E}[X]$, and $Z := t$ (note that a constant is a valid random variable). Then

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \mathbb{P}(\lambda(X - \mathbb{E}[X]) \geq \lambda t) = \mathbb{P}(e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}),$$

where we use the fact that $g_1(x) = \lambda x$, $\lambda \geq 0$ and $g_2(x) = e^x$ are non-decreasing and increasing, respectively. Now we can apply Markov's inequality, because $e^{\lambda(X - \mathbb{E}[X])}$ is a non-negative random variable:

$$\begin{aligned}\mathbb{P}(e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}) &\leq \frac{\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]}{e^{\lambda t}} \\ &= e^{-\lambda t - \lambda \mathbb{E}[X]} \mathbb{E}[e^{\lambda X}].\end{aligned}$$

Since λ was arbitrary and the bound holds *for all* non-negative λ, we can pick the λ that makes the bound smallest (i.e. most informative). That gives the final bound

$$\mathbb{P}(e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}) \leq \inf_{\lambda \geq 0} e^{-\lambda t - \lambda \mathbb{E}[X]} \mathbb{E}[e^{\lambda X}].$$

3. Prove that, if $X_1, \ldots, X_n$ are a sequence of independent random variables, then

$$\psi_{\sum_i \alpha_i X_i}(\lambda) = \prod_{i=1}^{n} \psi_{X_i}(\alpha_i \lambda),$$

where $\psi_Y(\lambda)$ is the moment-generating function of Y.

Solution: By definition, we write

$$\psi_{\sum_i \alpha_i X_i}(\lambda) = \mathbb{E}[e^{\sum_i \alpha_i X_i \lambda}].$$

By properties of the exponential function, we equivalently write this as

$$\mathbb{E}[e^{\sum_i \alpha_i X_i \lambda}] = \mathbb{E}[\prod_i e^{\alpha_i X_i \lambda}].$$

Now we use the fact that, if $Y_1, \ldots, Y_n$ are independent, then

$$\mathbb{E}[Y_1 Y_2 \ldots Y_n] = \prod_{i=1}^n \mathbb{E}[Y_i]. \tag{1}$$

Moreover, if $X_1, \ldots, X_n$ are independent, then $f_1(X_1), \ldots, f_n(X_n)$ are independent, for any sequence of functions $f_1, \ldots, f_n$. Therefore, we can conclude that

$$e^{\alpha_1 X_1 \lambda}, \ldots, e^{\alpha_n X_n \lambda}$$

are independent, and apply the rule given by equation (1):

$$\mathbb{E}[\prod_i e^{\alpha_i X_i \lambda}] = \prod_i \mathbb{E}[e^{\alpha_i X_i \lambda}] = \prod_i \psi_{X_i}(\alpha_i \lambda),$$

where in the last step we apply the definition of the MGF.

4. Let $X \sim N(\mu, \sigma^2)$. Prove that

$$\mathbb{P}(X - \mu \geq t) \leq e^{-t^2/2\sigma^2},$$

for all $t > 0$. Use the fact that the Gaussian moment-generating function is equal to $\psi(\lambda) = e^{\mu\lambda} e^{\frac{1}{2}\sigma^2\lambda^2}$ (it is a good exercise to convince yourself of this fact).

Solution: We apply the result of the previous part; in particular, we know

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \geq 0} e^{-\lambda t - \lambda\mu} \mathbb{E}[e^{\lambda X}] = \inf_{\lambda \geq 0} e^{-\lambda t} e^{-\lambda\mu} \mathbb{E}[e^{\lambda X}].$$

Now we apply the expression for the Gaussian moment-generating function to get

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \geq 0} e^{-\lambda t + \frac{1}{2}\sigma^2\lambda^2}.$$

This bound is smallest when the exponent is smallest (by monotonicity of the exponential function); therefore, we optimize the bound by setting the derivative of the exponent to 0:

$$-t + \sigma^2 \lambda^* = 0,$$

i.e. $\lambda^* = t/\sigma^2$. Plugging λ^* back into the bound gives

$$\mathbb{P}(X - \mu \geq t) \leq e^{-t^2/2\sigma^2},$$

as desired.

5. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be an iid sequence of Gaussians. Prove that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) \leq e^{-nt^2/2\sigma^2},$$

for all $t > 0$. What happens as $n \to \infty$?

Solution: We give two different solutions, using two different approaches.

Solution 1. We use the fact that linear combinations of Gaussian observations are also Gaussians: if $Y_i \sim N(\mu_i, \sigma_i^2)$ are independent, then

$$\sum_{i=1}^n \alpha_i Y_i \sim N(\sum_{i=1}^n \alpha_i \mu_i, \sum_{i=1}^n \alpha_i^2 \sigma_i^2).$$

Therefore, $\frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \frac{\sigma^2}{n})$. Now we just apply the result of the previous exercise to get the final inequality

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) \leq e^{-nt^2/2\sigma^2}.$$

Solution 2. As shown in exercise 3, $\psi_{\frac{1}{n}\sum_i X_i}(\lambda) = \prod_i \psi_{X_i}(\lambda/n)$. By applying the formula for the Gaussian MGF, we get $\psi_{\frac{1}{n}\sum_i X_i}(\lambda) = \left(e^{\mu\frac{\lambda}{n}} e^{\frac{1}{2}\sigma^2\frac{\lambda^2}{n^2}}\right)^n = e^{\mu\lambda} e^{\frac{1}{2}\sigma^2\frac{\lambda^2}{n}}$. We plug this into the Chernoff bound from exercise 2 to conclude

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq t\right) &\leq \inf_{\lambda \geq 0} e^{-\lambda t} e^{-\mathbb{E}[\frac{1}{n}\sum_i X_i]\lambda} \psi_{\frac{1}{n}\sum_i X_i}(\lambda) \\ &= \inf_{\lambda \geq 0} e^{-\lambda t} e^{-\mu\lambda} \psi_{\frac{1}{n}\sum_i X_i}(\lambda) = \inf_{\lambda \geq 0} e^{-\lambda t + \frac{1}{2}\sigma^2\frac{\lambda^2}{n}}.\end{aligned}$$

As in the previous part, we optimize the final expression over λ by setting the derivative of the exponent to 0, i.e. we set

$$-t + \sigma^2\lambda^*/n = 0.$$

This gives $\lambda^* = nt/\sigma^2$. Plugging this back into the bound completes the proof:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq t\right) \leq e^{-\frac{1}{\sigma^2}nt^2 + \frac{1}{2\sigma^2}nt^2} = e^{-nt^2/2\sigma^2}.$$

Now we take $n \to \infty$. We see that the bound tends to 0, i.e. $e^{-nt^2/2\sigma^2} \to 0$. What this says is that the probability of a sample average deviating from the mean by t, for *any* positive t, will tend to 0. This is essentially the Weak Law of Large Numbers, here proved for Gaussians.