High-Dimensional Statistics And The Johnson-Lindenstrauss Embedding

Daniel Raban

May 27, 2021

1 Overview of dimensionality

In the classical setup of statistics, we have N data points $\{u_1, \dots, u_N\} \subseteq \mathbb{R}^d$. Here, the parameter N represents the number of samples we draw our data from, while d represents the ambient dimension of the model, i.e. how many different features of the system we are measuring.

Example 1.1. We can poll $N = 1000$ UCLA students and ask them about their GPA and how much sleep they tend to get a night. Here, the number of features of the system is $d = 2$: GPA and hours of sleep.

In this regime, classical probability and statistics tell us that if we take enough samples (i.e. $N \to \infty$), our statistical estimates become precise.

Theorem 1.1 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \dots,$ be i.i.d. random variables with $\mathbb{E}[|X_1|] < \infty$. Then*

$$\frac{1}{N} \sum_{i=1}^{N} X_i \xrightarrow{a.s.} \mathbb{E}[X_1]$$

as $N \to \infty$.

Similarly, the Central Limit Theorem tells us what the distribution of $\frac{1}{N} \sum_{i=1}^{N} X_i$ looks like as $N \to \infty$.

More recent work in statistics has centered around a different situation: What if the number of samples N is much smaller than the number of features d?

Example 1.2. Suppose we are a biometrics company, and we want to do statistical analysis of the human genome. We receive the DNA of $N = 100$ people, but have $d = 21000$ genes.

In high-dimensional situations, we may not be able to use standard limit theorems. Here are two ways we can adapt to the issue:

1. Derive quantitative results which give explicit dependence on d and N.

2. Reduce the dimension without too much loss of information.

The Johnson-Lindenstrauss embedding achieves #2 using tools in the vein of #1. We will give basic probabilistic bounds in the form of **concentration inequalities** and use them to prove this theorem.

2 Markov's inequality and the Chernoff bound

Suppose you know the average value of a nonnegative random variable $X \geq 0$. If $\mathbb{E}[X] = 5$, for example, then the $\mathbb{P}(X \geq 100)$ has to be small; otherwise, these events would weight the expectation up too high, exceeding $\mathbb{E}[X]$. Markov's inequality says this in a quantitative form:

Theorem 2.1 (Markov's inequality)**.** *Let $X \geq 0$ be a nonnegative random variable. Then for any $a > 0$,*

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

Test your intuition with this by making a bigger. You get a tighter bound on the probability of the tail of the distribution.

Proof. Since $X \geq 0$,

$$\begin{aligned}\mathbb{E}[X] &\geq \mathbb{E}[X\mathbb{1}_{\{X \geq a\}}] \\ &\geq \mathbb{E}[a\mathbb{1}_{\{X \geq a\}}] \\ &= a(\cancel{0 \cdot \mathbb{P}(X < a)} + 1 \cdot \mathbb{P}(X \geq a)).\end{aligned}$$

Dividing both sides by a gives the inequality. □

Markov's inequality is not usually very tight, which you can expect by the fact that it only uses information given by the average value $\mathbb{E}[X]$. Here is a trick that often gives a tighter bound.

Theorem 2.2 (Chernoff bound)**.** *Let X be a random variable and $a > 0$. Then*

$$\mathbb{P}(X \geq a) \leq \inf_{t>0} e^{-at}\,\mathbb{E}[e^{tX}].$$

Remark 2.1. The quantity $\mathbb{E}[e^{tX}]$ is the **moment generating function** of X. Using the Maclaurin series for e^x, you can think of this as

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} \mathbb{E}[X^k]\frac{t^k}{k!},$$

which contains information about $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and more.

Remark 2.2. Also notice that the requirement $X \geq 0$ is gone. This is another strength of the Chernoff bound.

Proof. For any $t > 0$,

$$\begin{aligned} \mathbb{P}(X \geq a) &= \mathbb{P}(tX \geq ta) \\ &= \mathbb{P}(e^{tX} \geq e^{ta}) \end{aligned}$$

Using Markov's inequality,

$$\leq e^{-ta}\,\mathbb{E}[e^{tX}].$$

This bound holds for every $t > 0$, so we can take an inf on the right hand side to get the result. □

Let's see an example of this bound in action:

Example 2.1. Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable with mean μ and variance σ^2. We can compute $\mathbb{E}[e^{t(X-\mu)}] = e^{\sigma^2 t^2/2}$. Plugging this into the Chernoff bound gives

$$\mathbb{P}(X - \mu \geq a) \leq \inf_{t>0} \exp\left(\frac{\sigma^2 t^2}{2} - ta\right)$$

Using the fact that a parabola $at^2 + bt + c$ is minimized at $t = -b/(2a)$,

$$= \exp\left(-\frac{a^2}{2\sigma^2}\right)$$

We get the same thing if we apply the argument to $-X$, so adding the two cases together gives

$$\mathbb{P}(|X - \mu| \geq a) \leq 2\exp\left(-\frac{a^2}{2\sigma^2}\right).$$

3 Connecting the Chernoff bound to dimensionality

To see how the Chernoff bound relates to our original considerations, apply the previous example to an average of i.i.d. $N(\mu, \sigma)$ random variables: $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. $\overline{X}$ is a normally distributed random variable with mean μ and variance σ^2/N, so the Chernoff bound gives

Theorem 3.1 (Hoeffding's inequality for Gaussians)**.**

$$\mathbb{P}(|\overline{X} - \mu| \geq a) \leq 2\exp\left(-\frac{Na^2}{2\sigma^2}\right).$$

This type of inequality is called a **concentration inequality**. It tells us that the average $\overline{X}$ probabilistically concentrates around its mean at an exponential rate in N.

Using this same tool, we can gain information about high-dimensional data:

Lemma 3.1 (Bernstein's inequality for χ^2-RVs)**.** *Let $Z_1, Z_2, \ldots, Z_m$ be i.i.d. $N(0,1)$ random variables. Then for $0 < \delta < 1$,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i^2 - 1\right| \geq \delta\right) \leq 2\exp\left(-\frac{m\delta^2}{8}\right).$$

This says that Gaussian random vectors tend to concentrate around the sphere of radius $\sqrt{m}$.

Proof. The proof is the same as Hoeffding's inequality. We apply the Chernoff bound after calculating the moment generating function of $\frac{1}{m}\sum_{i=1}^{m} Z_i^2$. □

4 The Johnson-Lindenstrauss embedding

We are now able to prove the Johnson-Lindenstrauss theorem, which uses the above concentration inequalities

Theorem 4.1 (Johnson-Lindenstrauss)**.** *Consider a set of points $\{u_1, \ldots, u_N\} \subseteq \mathbb{R}^d$, and let $\varepsilon, \delta \in (0,1)$. For any m such that*

$$m \geq \frac{16}{\delta^2}\log\left(\frac{N}{\sqrt{\varepsilon}}\right),$$

if we define the random matrix

$$X := \frac{1}{\sqrt{m}} Z = \frac{1}{\sqrt{m}} \begin{bmatrix} — & Z_1 & — \\ & \vdots & \\ — & Z_m & — \end{bmatrix},$$

where Z is an $m \times d$ matrix with i.i.d. $N(0,1)$ entries, then

$$(1-\delta)\|u_i - u_j\|_2^2 \leq \|Xu_i - Xu_j\|_2^2 \leq (1+\delta)\|u_i - u_j\|_2^2 \qquad \forall i \neq j$$

with probability $\geq 1 - \varepsilon$.

In other words, given any distance distortion tolerance δ and probability tolerance ε, we can reduce the dimension of the data to $\sim \log N$.

Remark 4.1. Remarkably, the reduced dimension m only depends on the number of samples N; there is no dependence on the original dimension d. Moreover, it has the form of $\log N$, which grows even slower than N itself!

Remark 4.2. Not only does the theorem prove the existence of a linear map $\mathbb{R}^d \to \mathbb{R}^m$ embedding the data into a low-dimensional space, it also shows that "most" random linear maps will satisfy this property.

Proof. Fix $i \neq j$, and let $w = u_i - u_j$. First observe that the distortion bound is equivalent to

$$\left| \frac{\|Xw\|_2^2}{\|w\|_2^2} - 1 \right| \leq \delta.$$

Let $Y = \|Xw\|_2^2 / \|w\|_2^2$. Then

$$Y = \left\| X \frac{w}{\|w\|_2} \right\|_2^2 = \frac{1}{m} \sum_{i=1}^m \underbrace{\left\langle Z_i, \frac{w}{\|w\|_2} \right\rangle^2}_{\sim [N(0,1)]^2},$$

so $mY \sim \chi^2(m)$ is a chi-squared random variable. Using our concentration inequality lemma,

$$\mathbb{P}(|Y - 1| > \delta) \leq 2 \exp\left(-\frac{m\delta^2}{8} \right) \leq 2 \frac{\varepsilon}{N^2}.$$

Now, to treat all the pairs (i,j) together, we use a union bound:

$$\mathbb{P}\left(\left| \frac{\|X(u_i - u_j)\|_2^2}{\|u_i - u_j\|_2^2} - 1 \right| > \delta \text{ for some } i \neq j \right) \leq \sum_{i \neq j} \mathbb{P}\left(\left| \frac{\|X(u_i - u_j)\|_2^2}{\|u_i - u_j\|_2^2} - 1 \right| > \delta \right)$$

$$\begin{aligned} &\leq \binom{N}{2} 2 \frac{\varepsilon}{N^2} \\ &\leq \varepsilon. \end{aligned}$$

□

Remark 4.3. We do not need to use Gaussian random variables in this theorem. All these concentration inequalities can be applied more generally to **sub-Gaussian** random variables, which are any random variables where the tails of the distribution decay at least as fast as the Gaussian (e.g. bounded random variables). We can even apply the embedding with random variables which take the values ±1, each with probability 1/2 (this gives different constants for m).