Concentration of measure in probability and high-dimensional statistical learning

Lesson # 6

Guillaume Aubrun

Today we will look at high dimensions through the lens of geometry.
What does a high-dimensional space look like?

Placeholders marked **Proof** will be filled in class by writing on the slides.

We denote by $\mathrm{vol}(\cdot)$ the volume (=Lebesgue measure) in $\mathbf{R}^n$.
Given subsets A, B in $\mathbf{R}^n$ and $t \in \mathbf{R}$, define

$$tA = \{tx \ : \ x \in A \},$$

$$A + B = \{x + y \ : \ x \in A, y \in B \}.$$

We have $\mathrm{vol}(tA) = |t|^n \mathrm{vol}(A)$ by homogeneity. What about $\mathrm{vol}(A + B)$?

Theorem (The Brunn–Minkowski inequality)

If A, B are nonempty then $\mathrm{vol}(A + B)^{1/n} \geqslant \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}$.

A (globally) equivalent inequality is: for $t \in [0, 1]$,

$$\mathrm{vol}(tA + (1 - t)B) \geqslant \mathrm{vol}(A)^t \, \mathrm{vol}(B)^{1-t}.$$

Proof of equivalence.

There is a generalization which can be proved by induction on the dimension.

Theorem (The Prékopa–Leindler inequality)

Fix $t \in [0,1]$. If $f, g, h : \mathbf{R}^n \to [0,\infty)$ satisfy

$$h(tx + (1-t)y) \geqslant f(x)^t g(y)^{1-t}$$

then

$$\int h \geqslant \left(\int f\right)^t \left(\int g\right)^{1-t}.$$

The hypothesis is satisfied for $f = \mathbf{1}_A$, $g = \mathbf{1}_B$, $h = \mathbf{1}_{tA+(1-t)B}$, and the conclusion is precisely $\mathrm{vol}(tA + (1-t)B) \geqslant \mathrm{vol}(A)^t \,\mathrm{vol}(B)^{1-t}$.

Say that a probability measure μ on $\mathbf{R}^n$ is log-concave if it has a density f nd if $\log(f)$ is concave.
Examples: nondegenerate Gaussian measures, uniform measure on a convex set A with $0 < \text{vol}(A) < \infty$.

Proposition

A marginal of a log-*concave measure is* log-*concave.*

Proof.

Say that a r.v. X is log-concave if its distribution is log-concave.

Proposition

The sum of independent log-*concave r.v.s is* log-*concave.*

Proof.

Denote by B_n the Euclidean ball in $\mathbf{R}^n$

$$B_n = \{x \in \mathbf{R}^n \ : \ \sum x_i^2 \leqslant 1\}.$$

Let $K \subset \mathbf{R}^n$. The surface area of K can be defined as

$$a(K) = \limsup_{\varepsilon \to 0} \frac{\mathrm{vol}(K + \varepsilon B_n) - \mathrm{vol}(K)}{\varepsilon}.$$

Theorem (Isoperimetric inequality)

Let $K \subset \mathbf{R}^n$ with $\mathrm{vol}(K) > 0$ and $r > 0$ such that $\mathrm{vol}(K) = \mathrm{vol}(rB_n)$. Then $\mathrm{vol}(K + \varepsilon B_n) \geqslant \mathrm{vol}(rB_n + \varepsilon B_n)$ and therefore $a(K) \geqslant a(rB_n)$.

For given volume, surface area is minimized by Euclidean balls.

Proof.

Let σ be the uniform probability measure on the sphere $S^{n-1} = \partial B_n$. It can be defined for $A \subset S^{n-1}$ by

$$\sigma(A) = \frac{\mathrm{vol}_n(\{ta \ : \ t \in [0,1], a \in A\})}{\mathrm{vol}_n(B_n)}.$$

The measure σ is rotation-invariant.
There are two distances on S^{n-1}: the geodesic distance g and the Euclidean distance from $\mathbf{R}^n$, related by the formula

$$|x - y| = 2 \sin\left(\frac{g(x,y)}{2}\right).$$

Denote by $C(x,\theta)$ the spherical cap of center $x \in S^{n-1}$ and angle $\theta \in [0,\pi]$.

$$C(x,\theta) = \{y \in S^{n-1} \ : \ g(x,y) \leqslant \theta\}.$$

Let $V_n(\theta) = \sigma(C(x,\theta))$. We have $V_n(\pi - \theta) = 1 - V_n(\theta)$.

Crucial fact: for fixed $\theta < \pi/2$, $V_n(\theta)$ is exponentially small as $n \to \infty$.

We have $V_n(\theta) \leqslant \frac{1}{2}\sin(\theta)^{n-1}$
Proof

For $0 < \theta < \pi/2$, it can be shown that

$$V(\theta)^{1/n} \sim \sin(\theta)$$

as $n \to \infty$.

Let (M, d) be a compact metric space and $\varepsilon > 0$. A subset $N \subset M$ is

1. ε-dense (or a ε-net) if $\forall x \in M$, $\exists x_0 \in N$ with $d(x, x_0) \leqslant \varepsilon$.
2. ε-separated if $\forall x \neq y$ in N, $d(x, y) > \varepsilon$.

We then define

1. the **covering number** $N(M, \varepsilon)$ as the minimal cardinality of an ε-dense set,
2. the **packing number** $P(M, \varepsilon)$ as the minimal cardinality of an ε-separated set.

We have

$$P(M, 2\varepsilon) \leqslant N(M, \varepsilon) \leqslant P(M, \varepsilon).$$

Proof

Consider the metric space (S^{n-1}, g). We have the inequalities (**why?**)

$$\frac{1}{V_n(\varepsilon)} \leqslant N(S^{n-1}, \varepsilon) \leqslant P(S^{n-1}, \varepsilon) \leqslant \frac{1}{V_n(\varepsilon/2)}$$

therfore both $N(S^{n-1}, \varepsilon)$ and $N(S^{n-1}, \varepsilon)$ grow exponentially fast.
There is a precise answer for the covering number growth rate.

Theorem (Rogers)

For $0 < \varepsilon < \pi/2$ *we have* $\lim_{n\to\infty} \frac{1}{n} \log N(S^{n-1}, \varepsilon) = -\log \sin \varepsilon.$

The packing problem is much more complicated. It is **conjectured** that

$$\lim_{n\to\infty} \frac{1}{n} \log P(S^{n-1}, \varepsilon) = -\log \sin \varepsilon,$$

which means that you cannot pack caps more efficently than by the greedy algorithm. Connects to coding theory.

Rogers's random covering argument. We show that if $\varepsilon_2 < \varepsilon_1$ then

$$N(S^{n-1}, g, \varepsilon_1 + \varepsilon_2) \leqslant \left\lceil \frac{1}{V_n(\varepsilon_1)} \log\left(\frac{V_n(\varepsilon_1)}{V_n(\varepsilon_2)}\right) \right\rceil + \frac{1}{V_n(\varepsilon_1)}$$

which implies the theorem after some analysis.
Fix $N = \left\lceil \frac{1}{V_n(\varepsilon_1)} \log\left(\frac{V_n(\varepsilon_1)}{V_n(\varepsilon_2)}\right) \right\rceil$ and let $(x_i)_{1 \leqslant i \leqslant N}$ be i.i.d. random uniform points on S^{n-1}. The covered part $A = \bigcup C(x_i, \varepsilon)$ satisfies

$$\mathbf{E}\sigma(S^{n-1} \setminus A) = (1 - V_n(\varepsilon_1))^N \leqslant \exp(-NV_n(\varepsilon_1)) \leqslant \frac{V_n(\varepsilon_2)}{V_n(\varepsilon_1)}.$$

Choose x_i such that $\sigma(S^{n-1} \setminus A) \leqslant \frac{V_n(\varepsilon_2)}{V_n(\varepsilon_1)}$. Let $C(y_j, \varepsilon_2)_{1 \leqslant j \leqslant M}$ a maximal set of disjoint caps inside $S^{n-1} \setminus A$. By disjointedness, we have $MV_n(\varepsilon_2) \leqslant \sigma(S^{n-1} \setminus A)$ and therefore $M \leqslant \frac{1}{V_n(\varepsilon_1)}$. By maximality, we have

$$S^{n-1} \subset \bigcup_{i=1}^{N} C(x_i, \varepsilon_1 + \varepsilon_2) \cup \bigcup_{j=1}^{M} C(y_j, 2\varepsilon_2)$$

showing (using that $\varepsilon_2 \leqslant \varepsilon_1$) that $N(S^{n-1}, g, \varepsilon_1 + \varepsilon_2) \leqslant N + M$.

Let $E \subset S^{n-1}$ be an equator (e.g. $\{x_1 = 0\}$). The ε-neighbourhood of E

$$E_\varepsilon = \{x \in S^{n-1} \ : \ \exists y \in E \ : \ g(x,y) > \varepsilon\}$$

is the complement of two caps of angle $\pi/2 - \varepsilon$. Therefore

$$\sigma(S^{n-1} \setminus E_\varepsilon) = 2V_n(\pi/2 - \varepsilon) \leqslant \cos(\varepsilon)^{n-1} \leqslant \exp(-(n-1)\varepsilon^2/2).$$

Most of the mass on the sphere lies very close to an equator. This is even true simultaneously for $N \gg 1$ equators, as long as N is subexponential.

In other words, linear functions on the sphere are subGaussian r.v.s.

A much stronger statement is true: all Lipschitz functions are subGaussian r.v.s.

One can also prove an isoperimetric inequality on the sphere.

Theorem (Isoperimetric inequality on S^{n-1})

Let $A \subset S^{n-1}$ and C be a spherical cap such that $\sigma(A) = \sigma(C)$. Then for every $\varepsilon > 0$, we have $\sigma(A_\varepsilon) \geqslant \sigma(C_\varepsilon)$,

As a corollary, if $A \subset S^{n-1}$ satisfies $\sigma(A) = 1/2$, then $\sigma(S^{n-1} \setminus A_\varepsilon) \leqslant V(\pi/2 - \varepsilon) \leqslant \frac{1}{2} \exp(-(n-1)\varepsilon^2/2)$.

This estimate can also be deduced from Brunn–Minkowski inequality.

Proof

Corollary

Let $f : S^{n-1} \to \mathbf{R}$ a 1-Lipschitz function with median m. Then

$$\sigma(\{f \geqslant m + \varepsilon\}) \leqslant \frac{1}{2} \exp(-(n-1)\varepsilon^2/2),$$

$$\sigma(\{|f - m| \geqslant \varepsilon\}) \leqslant \exp(-(n-1)\varepsilon^2/2).$$

Proof

Computing the median is often not easy. But once we know a function concentrates, we can a posteriori replace the median by the mean.

Corollary

Let $f : S^{n-1} \to \mathbf{R}$ a 1-Lipschitz function with expectation $\mathbf{E}f$. Then

$$\sigma(\{|f - \mathbf{E}f| \geqslant \varepsilon\}) \leqslant C \exp(-cn\varepsilon^2).$$

Here C and c are absolute constants.

We compare the expectation and the mean

$$\begin{aligned} |\mathbf{E}f - m| \leqslant \mathbf{E}|f - m| &= \int_0^\infty \sigma(|f - m| \geqslant t)\,\mathrm{d}t \\ &\leqslant \int_0^\infty \exp(-(n-1)t^2)\,\mathrm{d}t \\ &= O(1/\sqrt{t}) \end{aligned}$$

so replacing m by $\mathbf{E}f$ only affects the constants.