../../Ik-Vault/Zettelkasten/Sub-Gaussian McDiarmid Inequality and Classification on the Sphere.md

Author

April 13, 2023

Abstract

We use the sub-Gaussian McDiarmid inequality to quantify the parametric error for binary classification on the sphere. We also include a proof of this inequality, which employs the entropy method.

longform

1 Introduction

project/toplevel Consider a binary linear classification problem with feature vectors $X_1, X_2, \ldots, X_m \overset{\text{iid}}{\sim} \text{Unif}(\mathbb{S}^{n-1})$ and corresponding labels $Y_i = \text{sign}(\langle w, X_i \rangle)$, where $w \in \mathbb{S}^{n-1}$ is fixed. The objective is to estimate w (or equivalently, learn the linear classifier $x \mapsto \langle w, \cdot \rangle$). Here, we study the statistical properties of the following quantity

$$\widetilde{w} := \frac{1}{m} \sum_{i=1}^{m} X_i Y_i, \tag{1}$$

which yields the estimator

$$\widehat{w} := \frac{\tilde{w}}{\|\mathbb{E}\tilde{w}\|_2}. \tag{2}$$

The deviation

$$\|\widetilde{w} - \mathbb{E}\widetilde{w}\|_2 = \left\| \frac{1}{m} \sum_{i=1}^{m} X_i Y_i - \mathbb{E}\widetilde{w} \right\|_2 = \left\| \frac{1}{m} \sum_{i=1}^{m} Z_i - \mathbb{E} Z_1 \right\|_2 ,$$

where $Z_i := X_i Y_i$ is now distributed uniformly on a half-sphere, is a well-behaved function of independent random variables. Hence, is amenable to concentration of measure principles. Here, we control the sub-Gaussian norm of this quantity using the *sub-Gaussian McDiarmid inequality* [**maurerConcentrationInequalitiesSubGaussian20**

Theorem 1.1 (Characterisation of Estimation Error)**.** *Suppose $n \in \mathbb{N}$. Then we have*

$$\mathbb{E}\left[\|\widehat{w} - w\|_2\right] \asymp \sqrt{\frac{n}{m}},$$

and

$$\left\|\|\widehat{w} - w\|_2 - \mathbb{E}\left[\|\widehat{w} - w\|_2\right]\right\|_{\psi_2} \lesssim \frac{1}{\sqrt{m}}.$$

2 Proof of ??

proof::We shall use the notation introduced in **??**. Without loss of generality, we may assume $w = e_1$. We compute

$$\begin{aligned}\|\mathbb{E}[\widetilde{w}]\|_2 &= \frac{\int_0^1 y(1-y^2)^{(n-3)/2} dy}{\int_0^1 (1-y^2)^{(n-3)/2} dy} \\ &= \frac{2\Gamma(n/2)}{\sqrt{\pi}(n-1)\Gamma((n-1)/2)} \\ &\asymp \frac{1}{\sqrt{n}}.\end{aligned}$$

We now compute the mean of the deviation $\|\widetilde{w} - \mathbb{E}[\widetilde{w}]\|_2$ as Further, we compute the deviation of $\|\widetilde{w} - \mathbb{E}[\widetilde{w}]\|_2$ from its mean as

$$\|\widetilde{w} - \mathbb{E}[\widetilde{w}]\|_2 = \left\|\frac{1}{m}\sum_{i=1}^m Z_i - \mathbb{E}[Z_1]\right\|_2 =: g(Z_1, \ldots, Z_m),$$

where $Z_1, \ldots, Z_m \overset{iid}{\sim} Unif(\mathbb{S}_+^{n-1})$. Let $i \in [m]$. Let $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_m \in \mathbb{S}^{n-1}$ be fixed. Note that $x \mapsto g(z_1, \ldots, z_{i-1}, x, z_{i+1}, \ldots, z_m)$ is a Lipschitz function from $\mathbb{S}^{n-1}$ to $\mathbb{R}$ with Lipschitz constant $1/m$ since

$$|g(z_1, \ldots, z_{i-1}, x, z_{i+1}, \ldots, z_m) - g(z_1, \ldots, z_{i-1}, x', z_{i+1}, \ldots, z_m)| \leq \frac{1}{m}\left\|x - x'\right\|_2$$

by the reverse triangle inequality. Then it follows from **??** that

$$\|g(z_1, \ldots, z_{i-1}, Z_i, z_{i+1}, \ldots, z_m) - \mathbb{E}[g(z_1, \ldots, z_{i-1}, Z_i, z_{i+1}, \ldots, z_m)]\|_{\psi_2} \lesssim \frac{1}{m\sqrt{n}}.$$

Finally, it follows from **??** that

$$\| \|\widetilde{w} - \mathbb{E}[\widetilde{w}]\|_2 - \mathbb{E}[\|\widetilde{w} - \mathbb{E}\,[\widetilde{w}]\|_2]\,\|_{\psi_2} = \|g(Z_1, \ldots, Z_m) - \mathbb{E}[g(Z_1, \ldots, Z_m)]\|_{\psi_2} \lesssim \frac{1}{\sqrt{mn}}.$$

The final result now follows by combining the above estimate with the estimates on $\|\mathbb{E}[\widetilde{w}]\|_2$ and $\|\widetilde{w} - \mathbb{E}[\widetilde{w}]\|_2$ proved before.

proof::Uses **??** together with **??**. From this we are done.

3 Proof of ??

proof::Let $Z := f(X_1, \ldots, X_m) - \mathbb{E}f(X_1, \ldots, X_m)$, and consider its log moment generating function $\psi(\lambda) = \log \mathbb{E}e^{\lambda Z}$, for $\lambda \in \mathbb{R}$. All we need to show is that $\psi(\lambda) \leq \frac{mK^2\lambda^2}{2}$ so that by definition of sub-Gaussian random variables, we are done. In fact, it suffices to prove this for $\lambda \geq 0$, because the case $\lambda < 0$ then follows by repeating the argument for $-Z$.

The proof we present here uses the *entropy method* (see [**boucheronConcentrationInequalitiesNonasym** [**wainwrightHighDimensionalStatisticsNonAsymptotic2019**]). definition:: For a non-negative random variable W , and the convex function $\phi(w) = w \log w$, define the ϕ -entropy of W as

$$Ent(W) = \mathbb{E}\,\phi(W) - \phi\,(\mathbb{E}W)$$

This quantity is well defined when both

$$W$$

and

$$\phi(W)$$

have finite expectations. Some basic properties of the entropy are provided in the appendix.

First, we bound the cumulant generating function by a function of the entropy.

Lemma 3.1 (Herbst's Argument)**.** *To bound the cumulant generating function it suffices to bound the entropy. Specifically,*

$$\psi(\lambda) = \lambda \int_0^\lambda \frac{Ent(e^{tZ})}{\varphi(t)t^2}\,dt.$$

Notice that if

$$Ent(e^{tZ}) \leq cmK^2t^2\varphi(t)$$

, then it follows that

$$\frac{\psi(\lambda)}{\lambda} \leq \int_0^\lambda cK^2\frac{t^2}{t^2}\,dt = cK^2\lambda^2$$

$$\implies \psi(\lambda) \le cmK^2\lambda^2.$$

and from this the statement follows. Therefore we need only bound the entropy.

Let us first define some notation for the conditioning of random variables: let $Z_i := \mathbb{E}[Z|X_1,\dots,X_i]$ and let $Z^{(i)} := \mathbb{E}[Z|X_1,\dots,X_{i-1},X_{i+1},\dots,X_m]$.

We employ a standard method to bound the entropy: the so-called tensorization of entropy**??**. lemma::Define $W = g(X_1,\dots,X_m)$, then

$$Ent(W) \le \mathbb{E}\sum_{i=1}^m Ent^{(i)}(W)$$

where $Ent^{(i)}(W) = \mathbb{E}^{(i)}[W\log W] - \mathbb{E}^{(i)}W\log\mathbb{E}^{(i)}W$.

From this lemma we find that

$$Ent(e^{-tZ}) \le \sum_{i=1}^m \mathbb{E}Ent^{(i)}(e^{-tZ}).$$

We now bound $Ent^{(i)}(e^{tZ})$. Fix $X_1,\dots,X_{i-1},X_{i+1},\dots,X_m$. Then notice that bounding $Ent^{(i)}(e^{(-tZ)})$ reduces to bounding the entropy of a one-dimensional sub-gaussian random variable. The sub-Gaussianity in this setting is given by assumptions in the statement.

We bound this in the following lemma. lemma::Let Z be a sub-Gaussian random variable with norm K . Then

$$Ent(e^{tZ}) \le CK^2t^2\varphi(t).$$

We can now complete the proof. We write

$$\begin{aligned}\frac{Ent(e^{tZ})}{\varphi(t)} &\le \frac{\mathbb{E}\sum_{i=1}^m Ent^{(i)}(e^{tZ})}{\varphi(t)}\\ &\le \mathbb{E}\sum_{i=1}^m \frac{CK^2t^2\mathbb{E}^{(i)}e^{Zt}}{\varphi(t)}\\ &\le CmK^2t^2\end{aligned}$$

and as argued above, we then have

$$\psi(\lambda) \le cmK^2\lambda^2$$

and therefore Z is sub-gaussian with norm $K\sqrt{m}$.

We now show the three lemmas that where used in the proof in order of appearance. proof::Observe that

$$\frac{\psi(\lambda)}{\lambda} = \int_0^\lambda \frac{d}{dt}\frac{\psi(t)}{t}\,dt = \int_0^\lambda \frac{Ent(e^{tZ})}{\varphi(t)t^2}\,dt.$$

Indeed, this follows from computing the derivative of $\frac{\psi(t)}{t}$.

$$\begin{aligned} \frac{d}{dt}\frac{\psi(t)}{t} &= \frac{\psi'(t)t - \psi(t)}{t^2} \\ &= \frac{\left(\frac{\mathbb{E}[Ze^{tZ}]}{\varphi(t)}t - \frac{\varphi(t)\psi(t)}{\varphi(t)}\right)}{t^2} \\ &= \frac{\mathbb{E}[tZe^{tZ}] - \mathbb{E}[e^{tZ}]\log\mathbb{E}[e^{tZ}]}{t\varphi(t)} \\ &= \frac{Ent(e^{tZ})}{t^2\varphi(t)} \end{aligned}$$

To show the **??**, we will need the following result. lemma::Given a random variable Y ,

$$Ent(Y) = \sup_{U:\mathbb{E}e^U \le 1} \mathbb{E}[UY].$$

and furthermore, if we have a random variable U such that $\mathbb{E}[UY] \le Ent(Y)$ for any r.v. Y , then $\mathbb{E}e^U \le 1$.

proof::Consider

$$Ent_{e^U P}[e^{-U}Y]$$

and notice we can compute that

$$Ent_{e^U P}[e^{-U}Y] = Ent(Y) - \mathbb{E}[UY].$$

Then the proof follows from the fact that

$$Ent_{e^U P}[e^{-U}Y] \ge 0$$

with equality when the random variable is constant, i.e. when $e^{-U} = \frac{\mathbb{E}y}{Y}$ which yields a valid U , and so the inequality is attained for some U .

proof::First we need lemma::Given a random variable Y ,

$$Ent(Y) = \sup_{U:\mathbb{E}e^U \le 1} \mathbb{E}[UY].$$

and furthermore, if we have a random variable U such that $\mathbb{E}[UY] \le Ent(Y)$ for any r.v. Y , then $\mathbb{E}e^U \le 1$.

Define $W_i := \mathbb{E}[W|X_1, \ldots, X_i]$. Then

$$
\begin{aligned}
Ent(W) &= \mathbb{E}\left[W(\log W - \log \mathbb{E}W)\right] \\
&= \mathbb{E}\left[W \sum_{i=1}^{m} (\log W_i - \log \mathbb{E}^{(i)} W_i)\right] \\
&= \mathbb{E} \sum_{i=1}^{m} \mathbb{E}^{(i)}[W(\log W_i - \log \mathbb{E}^{(i)} W_i)] \\
&\leq \mathbb{E} \sum_{i=1}^{m} Ent^{(i)}(W)
\end{aligned}
$$

proof::Suppose $\mathbb{E}Z = 0$ first, and let $P = \mathrm{Law}(Z)$. For $t \in \mathbb{R}$, consider the exponentially tilted measure $dP^{(t)} = \frac{e^{tZ}}{\mathbb{E}e^{tZ}} dP$. Then,

$$
\begin{aligned}
\frac{Ent(e^{tz})}{\varphi(t)} &= \frac{\mathbb{E}[e^{tZ} tZ]}{\varphi(t)} - \frac{\varphi(t) \log \mathbb{E}e^{tZ}}{\varphi(t)} \\
&= \mathbb{E}_{P^{(t)}}[tZ] - \psi(t) \\
&= \mathbb{E}_{P^{(t)}} \log e^{tZ} - \psi(t) \\
&\leq \log \mathbb{E}_{P^{(t)}}[e^{tZ}] - \psi(t) \\
&= \log \mathbb{E}e^{2tZ} - 2\psi(t) \\
&\leq \log \mathbb{E}e^{2tZ} \\
&\leq CK^2 t^2
\end{aligned}
$$

This also holds for random variables of mean non-zero. Indeed,

$$
\frac{Ent(e^{tZ+C})}{\mathbb{E}[e^{tZ+C}]} = \frac{\mathbb{E}[e^C]}{\mathbb{E}[e^C]} \frac{Ent(e^{tZ})}{\varphi(t)} \leq CK^2 t^2.
$$

We were able to remove the negative term since $\psi(t) \geq 0$ because one can notice that for t ≥ 0 ,

$$
-\psi(t) = -\log(\mathbb{E}e^{Zt}) \leq -\mathbb{E}[Zt] \leq -t \leq 0.
$$

4 Appendix

Here, we prove some basic properties of the entropy functional, and state its relationship with the Kullback-Leibler divergence. theorem::Let $W \geq 0$ be any random variable with $\mathbb{E}W < \infty$ and $\mathbb{E}\,\phi(W) < \infty$.

- NOT IMPLEMENTED: VectorAny
- NOT IMPLEMENTED: VectorAny

- NOT IMPLEMENTED: VectorAny

proof::(1) follows directly from the convexity of $\phi(w) = w \log w$ by Jensen's inequality. To prove (2), suppose $W = c$ almost surely. Then $\mathbb{E}\phi(W) = \mathbb{E}\phi(c) = \phi(c) = \phi(\mathbb{E}W)$, giving that $Ent(W) = 0$. (3) follows from direct computation as

$$\begin{aligned} Ent(aW) &= \mathbb{E}[aW \log aW] - \mathbb{E}(aW)\log(\mathbb{E}aW) \\ &= a\left\{\mathbb{E}[W \log W] + (\mathbb{E}W)\log a - \mathbb{E}(W)\log(\mathbb{E}W) - (\mathbb{E}W)\log a\right\} \\ &= a\, Ent(W). \end{aligned}$$

The entropy functional is related to the usual Kullback-Leibler divergence of appropriately constructed measures. Suppose $P = \mathrm{Law}(X_1, \ldots, X_m)$ and let $W^{(\lambda)} = e^{\lambda g(X_1,\ldots,X_m)}$ for $\lambda \in \mathbb{R}$.

5 Statement

Define the corresponding *tilted* measure defined by the density

$$\frac{dP^{(\lambda)}}{dP} = \frac{e^{\lambda g(X_1,\ldots,X_m)}}{\mathbb{E}\, e^{\lambda g(X_1,\ldots,X_m)}}.$$

Then,

$$\begin{aligned} D_{\mathrm{KL}}(P^{(\lambda)}||P) &= \mathbb{E}\left[\frac{dP^{(\lambda)}}{dP}\log\frac{dP^{(\lambda)}}{dP}\right] \\ &= \mathbb{E}\left[\frac{dP^{(\lambda)}}{dP}\log\frac{dP^{(\lambda)}}{dP}\right] - \mathbb{E}\left[\frac{dP^{(\lambda)}}{dP}\right]\log\mathbb{E}\left[\frac{dP^{(\lambda)}}{dP}\right] \\ &= Ent\left(\frac{dP^{(\lambda)}}{dP}\right) \\ &= \frac{1}{\mathbb{E}\, e^{\lambda g(X_1,\ldots,X_m)}}\, Ent\left(e^{\lambda g(X_1,\ldots,X_m)}\right). \end{aligned}$$

where the second equality is due to the fact that $\mathbb{E}\left[\frac{dP^{(\lambda)}}{dP}\right] = 1$, and the last inequality follows from the positive homogeneity of entropy.