High-d (Heidi) Swiss Army Knife

Luis Oala

August 9, 2020

Outline

Justin Gilmer: High-d Error Rates

Eric Nalisnick: Degenerate OOD detection in High-d

vd Oord and Theis: How Likelihoods Can Break in High-d

Charu Aggarwal: High-d Metric Surprises

Justin Gilmer: High-d Error Rates

Justin Gilmer: High-d Error Rates

Setup[1]

- **error set** E: set of points in the input space on which the classifier makes an incorrect prediction
- **corruption robustness** $\mathbb{P}_{x \sim q}[x \notin E]$: probability that a random sample from the q is not an error, under a given corrupted image distribution q.
- **adversarial robustness** $\mathbb{P}_{x \sim p}[d(x, E) > \epsilon]$: probability that a random sample from p is not within distance ϵ of some point in the error set, where metric on the input space $d(x, E)$ denotes the distance from clean input x to the nearest point in E (also based on work by [2])
- **error rate** μ: $\mathbb{E}_{x \sim \mathcal{N}(x_0; \sigma^2 I)}[x \in E]$, with some clean image x_0 and the Gaussian distribution $\mathcal{N}(x_0; \sigma^2 I)$
- $\sigma(x_0, \mu)$: For a fixed μ, the σ for which the error rate is μ

[1] https://slideslive.com/38930579/why-adversarial-examples-feel-like-bugs

Justin Gilmer: High-d Error Rates (cont'd)

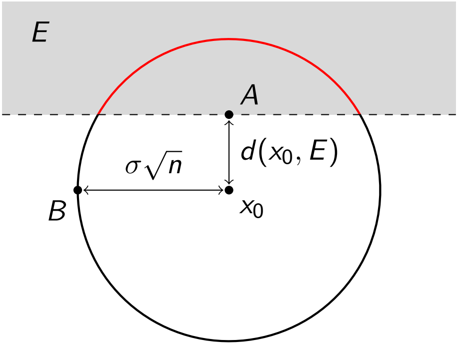


Figure 1: Sphere cutting illustration by [3]

Justin Gilmer: High-d Error Rates (cont'd)

Analysis

Letting d denote l_2 distance, we have

$$d(x_0, E) = -\sigma(x_0, \mu)\Phi^{-1}(\mu), \qquad (1)$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp(-x^2/2)dx$$

is the cdf of the univariate standard normal distribution. (Note that $\Phi^{-1}(\mu)$ is negative when $\mu < \frac{1}{2}$.)

Justin Gilmer: High-d Error Rates (cont'd)

Observations

- Equation 1 does not depend on n

n	$\sigma\sqrt{n}$	$d(x_0, E)$
3	0.17	0.23
150,528 (ImageNet)	38.8	0.23

Table 1: Linear model - distance of typical corrupted input ($\sigma\sqrt{n}$) and distance of nearest error ($d(x_0, E)$) under varying input dimension n

- $\frac{d(x_0,E)}{\sigma\sqrt{n}}$

Justin Gilmer: High-d Error Rates (cont'd)

Geometric Interpretation of Equation 1: Gaussian Annulus Theorem

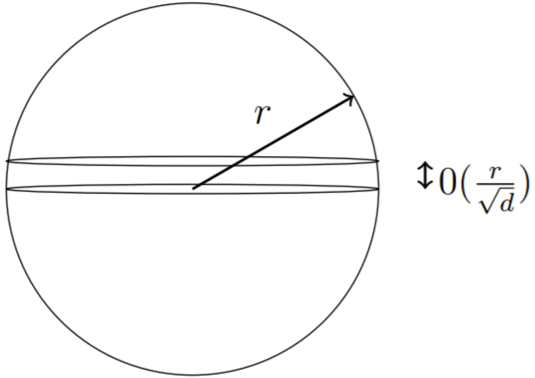


Figure 2: Equator concentration of spherical volume and surface area [2]

[2] https://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/

Eric Nalisnick: Degenerate OOD detection in High-d

Eric Nalisnick: Degenerate OOD detection in High-d

Decision rule:

$$p(\mathbf{X}^* \mid \text{IN}) > \frac{p(\mathbf{X}^* \mid \text{OUT})\, p(\text{OUT})}{p(\text{IN})}$$

(a)

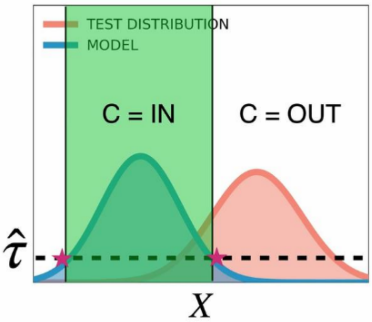
Decision rule:

$$\overset{\text{DGM}}{q(\mathbf{X}^*)} > \underbrace{\frac{\text{UNIFORM}(\mathbf{X}^*)\, p(\text{OUT})}{p(\text{IN})}}_{\hat{\tau}}$$

Implies classifier is just a threshold on the density function:

$$q(\mathbf{X}^*) > \hat{\tau}$$

(b)



(c)

Eric Nalisnick: Degenerate OOD detection in High-d (cont'd)

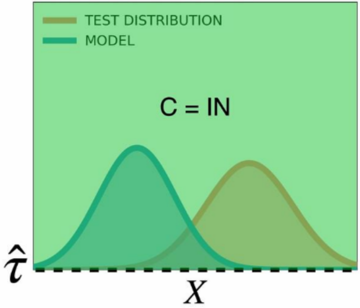
PROBLEM: In high-dimensions, the uniform OOD model becomes degenerate.

$$\text{UNIFORM}(\mathbf{x}) = \frac{1}{(b-a)^D} \rightarrow 0 \quad \text{as} \quad D \rightarrow \infty$$

Which leads to the degenerate threshold:

$$q(\mathbf{X}^*) > \text{UNIFORM}(\mathbf{X}^*) \frac{p(\text{OUT})}{p(\text{IN})} = 0$$

(d)



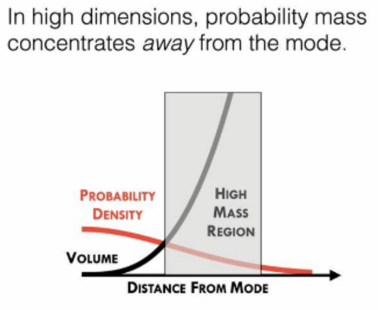
(e)

Note: All visualizations by Eric Nalisnick, check his talk at `https://icml.cc/virtual/2020/workshop/5742`

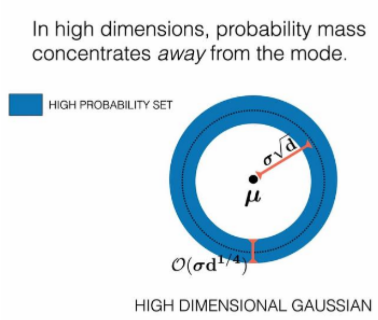
Eric Nalisnick: Degenerate OOD detection in High-d (cont'd)

Bonus Heidi

$m = V \times \rho$



(f)



(g)

vd Oord and Theis: How Likelihoods Can Break in High-d

vd Oord and Theis: How Likelihoods Can Break in High-d

[4, 5]

Great log-likelihood and poor samples

- p: density of a model for d dimensional data $\mathbf{x}$ which performs arbitrarily well with respect to average log-likelihood
- q: corresponds to some bad model (e.g., white noise)
- Then samples generated by the mixture model $0.01p(\mathbf{x}) + 0.99q(\mathbf{x})$ will come from the poor model 99% of the time
- Yet the log-likelihood per pixel will hardly change if d is large: $\log\left[0.01p(\mathbf{x}) + 0.99q(\mathbf{x})\right] \geq \log\left[0.01p(\mathbf{x})\right] = \log p(\mathbf{x}) - \log 100$ For high-dimensional data, $\log p(\mathbf{x})$ will be proportional to d while $\log 100$ stays constant.

Charu Aggarwal: High-d Metric Surprises

Charu Aggarwal: High-d Metric Surprises

[1] https://bib.dbvis.de/uploadedFiles/155.pdf

References I

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. "On the surprising behavior of distance metrics in high dimensional space". In: *International conference on database theory*. Springer. 2001, pp. 420–434.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. "Robustness of classifiers: from adversarial to random noise". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1632–1640.

Justin Gilmer et al. "Adversarial examples are a natural consequence of test error in noise". In: *International Conference on Machine Learning*. 2019, pp. 2280–2289.

Aäron van den Oord and Joni Dambre. "Locally-connected transformations for deep gmms". In: *International Conference on Machine Learning (ICML): Deep learning Workshop*. 2015, pp. 1–8.

References II

Lucas Theis, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models". In: *arXiv preprint arXiv:1511.01844* (2015).