Unit III: Anomaly Detection

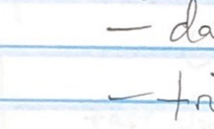
- General, difficult ML problem: how to detect data that is anomalous?
  ↓
  usually "less-than-supervised" — anomalous means don't know what you're looking for
  → lack labels

- Many scientific applications of anomaly detection
  - data quality monitoring
  - triggering
  - new physics searches — "model-agnostic" or "model-indep"
    ↳ vs "model-specific" → 99.99% of all current LHC searches!
    ↓
    what if we haven't found NP because we haven't searched in the right places yet?!
  - data-driven discovery

- Two main classes of anomaly detection:

1. Outlier detection

← outliers



looking for out of sample or out of dist'n (OOD) data

Outlier detection

- A clustering problem → many methods for unsupervised clustering
- also, a density estimation problem

↓

e.g. k means

DBSCAN

⋮

group data x_i into k clusters w/ means $\vec{\mu}_a$

must specify in advance!

$$\min \sum_{a=1}^{k} \sum_{x \in C_a} \|x - \vec{\mu}_a\|^2$$

fast algorithms exist for partitioning data & minimizing wrt $\vec{\mu}_a$.

main drawbacks → must specify metric in advance

⟹ what if the space in which data clusters nicely is hidden? (only "see" anomalies in latent space?)

---

density estimation: what data for which $p(x) = 0$? truly OOD

But that is self-contradictory

really mean $p_{bg}(x) = 0$?

But how will one know $p_{bg}(x)$? Not fully datadriven? Have pure sample of only bg?

In practice can never be sure $p_{bg}(x)=0$

(finite resolution, measurement error, quantum mechanics...)

↓

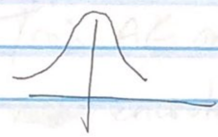
So really mean $p_{bg}(x) \ll 1$?

↓

But this is not coordinate inv't!

→ can change $p_{bg}(x)$ to anything w/ $x \to y$

extreme example:



$p(x) = Ne^{-x^2/2}$

$x = 5$ is very anomalous?

↓ CDF $\quad y = \int_{-\infty}^{x} dx' \, N e^{-x'^2/2}$

$p(y) = 1$ (uniform) $\quad y \in [0,1]$

So no points are outliers?!

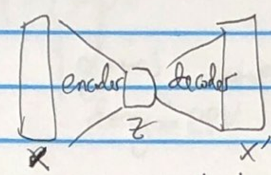
This issue is generally ignored in the density estimation based literature...

Hope that there is a preferred coord frame "physically meaningful"

General Methods for finding latent space clustering:

— Autoencoders (and variational AEs)

general idea:



$L = \|x - x'\|^2$ "reconstruction error"

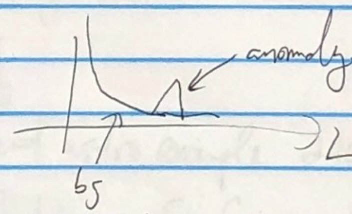
Learn to map data back to itself through compressed latent space

"information bottleneck"

↳ can't learn identity map $x' = x$

"one class classification"
- Train AE on "normal events" → learn to reconstruct well
- encounter rare, anomalous event → don't reconstruct well

→ can use L itself as anomaly score!



Do you really need sample of "normals" to train on? Then not fully unsupervised...

→ in practice AE works on bg + small amount of anomaly! finite model capacity → still mostly learns to reconstruct bg well.

Example: Farina, Nakai & DS 1808.08992

1st appl. of AE anomaly detection to HEP!

(see also Heimel et al 1808.08979)

bg = QCD jet images

sig = { tops
gluinos
decay via RPV }

challenges:
- bg estimation?
  - double AE idea
    2111.06417
- complexity bias?
  train on tops, doesn't detect QCD...

↓

"normalized AE"
more like density estimation
2105.05735 (ML)
2206.14825 (HEP)

VAEs can also be used for anomaly detection
- can use reco error like AE
- or can look for outliers in latent space!

Many examples in literature
(eg VAE on MNIST missing one digit)

↓

recent astro example 2103.12102

Rubin proof-of-concept → can detect rare transients as anomalous using latent space clustering