Dive into Deep Learning 4.9

Environment and Distribution Shift

MinDong Sung, M.D.

DHLab, Yonsei University College of Medicine

2022-07-08

Introduction

- Where data come from in the first place or
- What we plan to ultimately do with the outputs from our models.

Introduction

- Sometimes models appear to perform marvelously as measured by test set accuracy but fail catastrophically in deployment when the distribution of data suddenly shifts.
- More insidiously, sometimes the very deployment of a model can be the catalyst that perturbs the data distribution.

Example

- A model to predict who will repay vs. default on a loan
    - Oxfords indicate repayment,
    - Sneakers indicate default
- For starters, as soon as we began making decisions based on footwear, customers would catch on and change their behavior.
- Before long, all applicants would be wearing Oxfords, without any coinciding improvement in credit-worthiness.

Introduction

- Aim
  - Expose some common concerns, and
  - Stimulate the critical thinking required to detect these situations early, mitigate damage,
  - Use machine learning responsibly.
- Solutions
  - Simple (ask for the "right" data),
  - Technically difficult (implement a reinforcement learning system),
  - others require that we step outside the realm of statistical prediction altogether and grapple with difficult philosophical questions concerning the ethical application of algorithms.

Type of Distribution Shift

- Passive prediction setting
  - data distributions might shift
  - what might be done to salvage model performance
- Classic setup
  - Training data ~ $p_S(\mathbf{x}, y)$
  - Test data ~ $p_T(\mathbf{x}, y)$
- Absent any assumptions on how p_S and p_T relate to each other, learning a robust classifier is impossible.
- Under some restricted assumptions on the ways our data might change in the future, principled algorithms can detect shift and sometimes even adapt on the fly, improving on the accuracy of the original classifier.

Covariate Shift

- We assume that:
  - The distribution of inputs $P(\mathbf{x})$ may change over time,
  - The labeling function, i.e., the conditional distribution $P(y \mid \mathbf{x})$ does not change.
- The natural assumption to invoke in settings where we believe that $\mathbf{x}$ causes y.
- The problem arises due to a shift in the distribution of the covariates (features)

Covariate Shift

Train dataset

cat

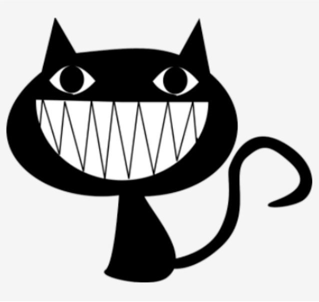
cat

dog

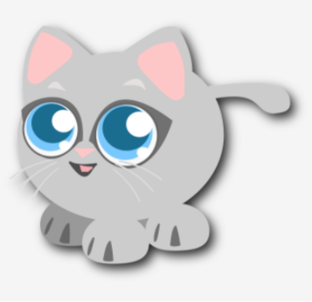
dog

Test dataset

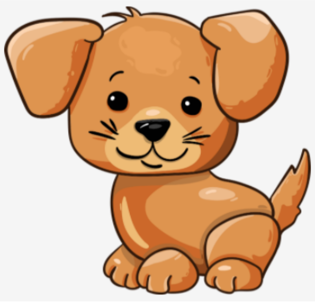
cat



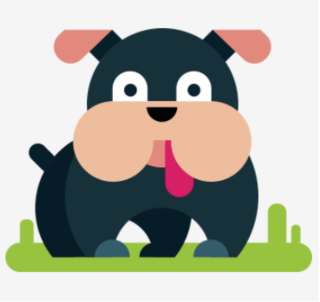
cat



dog



dog



Label Shift

- We assume that:
  - The distribution of labels $P(\mathbf{y})$ may can change
  - The class-conditional distribution $P(\mathbf{x} \mid y)$ remains fixed across domains.
- a reasonable assumption to make when we believe that y causes $\mathbf{x}$.

Example

- Predict diagnoses given their symptoms (or other manifestations), even as the relative prevalence of diagnoses are changing over time.
- Label shift is the appropriate assumption here because diseases cause symptoms.

Label Shift + Covariate Shift

- When the label is deterministic, the covariate shift assumption will be satisfied

Concept Shift

- The very definitions of labels can change.
- The distribution $P(y \mid \mathbf{x})$ might be different depending on our envrionment
- We might hope to exploit knowledge that shift only takes place gradually either in a temporal or geographic sense.

Example of Distribution Shift

Medical Diagnostics

- Due to their sampling procedure, we could expect to encounter extreme covariate shift.
- Moreover, this case was unlikely to be correctable via conventional methods.
- In short, they wasted a significant sum of money.

Self-Driving Cars

- Toadside detector
  - test in synthetic data from a game engine Vs. real world
- Detect tanks in the forest
  - the first set of pictures was taken in the early morning, the second set at noon.

Example of Distribution Shift

Nonstationary Distributions

- the distribution changes slowly (also known as *nonstationary distribution*)
- the model is not updated adequately.

Example

- A computational advertising model: newly appeared model
- Spam filter: new method detect
- Product recommendation system: Christmas?

More Anecdotes

- Face detector: with close up dataset
- Web search engine for the US market and want to deploy it in the UK
- Train in balanced class -> apply to unbalanced label distribution

Correction of Distribution Shift

Empirical Risk and Risk

- training data: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- model: f
- loss function: l
- minimizing loss: $\text{minimize}_f \frac{1}{n} \sum_{i=1}^n l(f(\mathbf{x}_i), y_i)$,
- trud distribution of data: $p(\mathbf{x}, y)$
- empirical risk

$$E_{p(\mathbf{x},y)}[l(f(\mathbf{x}), y)] = \int\int l(f(\mathbf{x}), y)p(\mathbf{x}, y)\; d\mathbf{x}dy.$$

c.f. $E[X] = \sum_{i=1}^n x_i p(x_i)$

- However, in practice we typically cannot obtain the entire population of data.
- *empirical risk minimization* ~ minimizing the risk

Covariate Shift Correction

- labeled data: $(\mathbf{x}_i, y_i)$
- source data: $q(\mathbf{x})$ | target data: $p(\mathbf{x})$
- $$\mathbf{x}_i \in q(\mathbf{x})$$
- covariate shift assumption: $p(y \mid \mathbf{x}) = q(y \mid \mathbf{x})$

$$\int\int l(f(\mathbf{x}), y) p(y \mid \mathbf{x}) p(\mathbf{x}) \; d\mathbf{x} dy = \int\int l(f(\mathbf{x}), y) q(y \mid \mathbf{x}) q(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} \; d\mathbf{x} dy.$$

- the ratio of the probability: $\beta_i \stackrel{\mathrm{def}}{=} \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$.

$$\underset{f}{\operatorname{minimize}} \frac{1}{n} \sum_{i=1}^{n} \beta_i l(f(\mathbf{x}_i), y_i).$$

- However, we don't know the ratio.

- Note that for any such approach, we need samples drawn from both distributions---the "true" p

- e.g. Logistic regression

- we have an equal number of instances from both distributions $p(\mathbf{x})$ and $q(\mathbf{x})$
- Now denote by z labels that are 1 for data drawn from p and -1 for data drawn from q.

$$P(z=1 \mid \mathbf{x}) = \frac{p(\mathbf{x})}{p(\mathbf{x})+q(\mathbf{x})} \text{ and hence } \frac{P(z=1 \mid \mathbf{x})}{P(z=-1 \mid \mathbf{x})} = \frac{p(\mathbf{x})}{q(\mathbf{x})}.$$

- Thus, if we use a logistic regression approach, where $P(z=1 \mid \mathbf{x}) = \frac{1}{1+\exp(-h(\mathbf{x}))}$

$$\beta_i = \frac{1/(1+\exp(-h(\mathbf{x}_i)))}{\exp(-h(\mathbf{x}_i))/(1+\exp(-h(\mathbf{x}_i)))} = \exp(h(\mathbf{x}_i)).$$

- Solve two problems
  - Distinguish between data drawn from both distributions
  - Weighted empirical risk minimization problem

Algorithm

- a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- an unlabeled test set $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$
- $\mathbf{x}_i$ for all $1 \leq i \leq n$ are drawn from some source distribution.
- $\mathbf{u}_i$ for all $1 \leq i \leq m$ are drawn from the target distribution.

1. Generate a binary-classification training set: $\{(\mathbf{x}_1, -1), \ldots, (\mathbf{x}_n, -1), (\mathbf{u}_1, 1), \ldots, (\mathbf{u}_m, 1)\}$.
2. Train a binary classifier using logistic regression to get function h.
3. Weigh training data using $\beta_i = \exp(h(\mathbf{x}_i))$ or better $\beta_i = \min(\exp(h(\mathbf{x}_i)), c)$ for some constant c.
4. Use weights β_i for training on $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

- We need that each data example in the target (e.g., test time) distribution had nonzero probability of occurring at training time.

Label Shift Correction

- q: the source distribution (e.g., training time)
- p: the target distribution (e.g., test time)
- $q(\mathbf{x} \mid y) = p(\mathbf{x} \mid y)$

$$\int\int l(f(\mathbf{x}), y) p(\mathbf{x} \mid y) p(y) \; d\mathbf{x} dy = \int\int l(f(\mathbf{x}), y) q(\mathbf{x} \mid y) q(y) \frac{p(y)}{q(y)} \; d\mathbf{x} dy.$$

- label likelihood ratios:

$$\beta_i \stackrel{\mathrm{def}}{=} \frac{p(y_i)}{q(y_i)}.$$

- Confusion Matrix $\mathbf{C}$: $k \times k$ matrix

$$\mathbf{C} p(\mathbf{y}) = \mu(\hat{\mathbf{y}}),$$

- Average all of our models predictions at test time together, yielding the mean model outputs $\mu(\hat{\mathbf{y}}) \in \mathbb{R}^k$, whose i^{th} element $\mu(\hat{y}_i)$ is the fraction of total predictions on the test set where our model predicted i.
- $\sum_{j=1}^{k} c_{ij} p(y_j) = \mu(\hat{y}_i)$ holds for all $1 \leq i \leq k$
- If our classifier is sufficiently accurate to begin with, then the confusion matrix $\mathbf{C}$ will be invertible, and we get a solution $p(\mathbf{y}) = \mathbf{C}^{-1}\mu(\hat{\mathbf{y}})$

Concept Shift Correction

- Concept shift is much harder to fix in a principled manner.
- e.g. Suddenly the problem changes

- It will be unreasonable to assume that we can do much better than just collecting new labels and training from scratch.

- Instead, what usually happens is that the task keeps on changing slowly.

- We use the existing network weights and simply perform a few update steps with the new data rather than training from scratch.

A Taxonomy of Learning Problems

Batch learning

- Learning once

Online learning

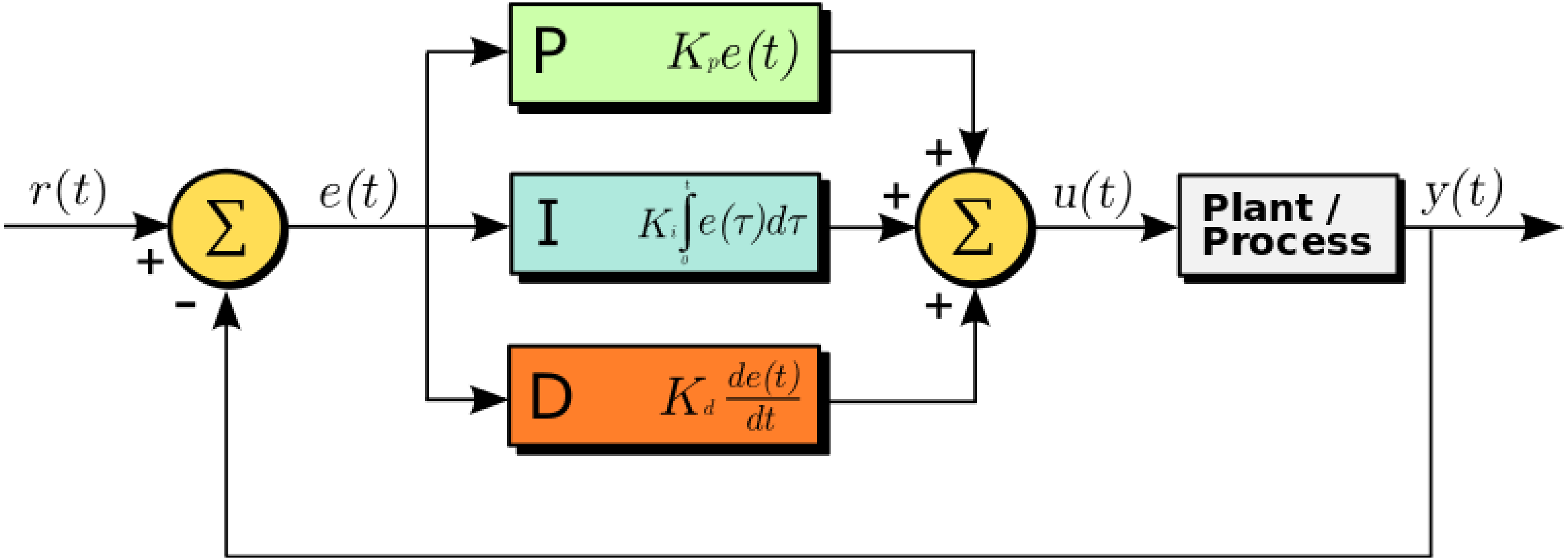
$$\text{model } f_t \longrightarrow \text{data } \mathbf{x}_t \longrightarrow \text{estimate } f_t(\mathbf{x}_t) \longrightarrow \text{observation } y_t \longrightarrow \text{loss } l(y_t, f_t(\mathbf{x}_t)) \\ \longrightarrow \text{model } f_{t+1}$$

Bandits

- We only have a finite number of arms that we can pull
- classic reinforcement learning problem

Control (Theory)

- environment remembers what we did



Reinforcement Learning

- In the more general case of an environment with memory, we may encounter situations where the environment is trying to cooperate with us (cooperative games, in particular for non-zero-sum games), or others where the environment will try to win.

Considering the Environment

- One key distinction between the different situations above is that the same strategy that might have worked throughout in the case of a stationary environment, might not work throughout when the environment can adapt.

Fairness, Accountability, and Transparency in Machine Learning

- When you deploy machine learning systems you are not merely optimizing a predictive model---you are typically providing a tool that will be used to (partially or fully) automate decisions
- The leap from considering predictions to decisions raises not only new technical questions, but also a slew of ethical questions that must be carefully considered
- Moreover, once we contemplate decision-making systems, we must step back and reconsider how we evaluate our technology.
- We also want to be careful about how prediction systems can lead to feedback loops.
- The various mechanisms by which a model's predictions become coupled to its training data are unaccounted for in the modeling process(runaway feedback loops)
- Predictive algorithms now play an outsize role in mediating the dissemination of information.