Model drift: when predictions become less accurate

Nikolay Manchev

February 26, 2020

Domino Data Lab

Housekeeping

- This is our first meetup event (Yay!)
- I'll aim to do one every 4 weeks (last Thursday of each month)
- Next event is on 27th March
- I need your help
    - Speakers
    - Topics
- All content will be on GitHub (under CC)
- To get in touch with me use **@nikolaymanchev**

Incremental Learning

Assumption: Training data contains all necessary information to learn the underlying function

Not always the case

- Weather forecast
- Epidemiological studies
- Spam detection

Sometimes data arrives in sequential fashion - we want to use all data available at time step t to predict what happens at $t + 1$

Incremental Learning

Formal definition [MTP09]

A learning algorithm is incremental if

- for a sequence of training instances produces a sequence of hypotheses
- the current hypothesis describes all data seen thus far
- the current hypothesis depends only on previous hypotheses and the current training data

Stability-plasticity dilemma [MBB13]

For any artificial or biological neural system that learns

- A system must stay stable and unchanged to irrelevant events
- The system must be plastic to new data

Stability-plasticity defines a scale

- Stability-end — batch learning
- Plasticity-end — on-line learning

Dataset types

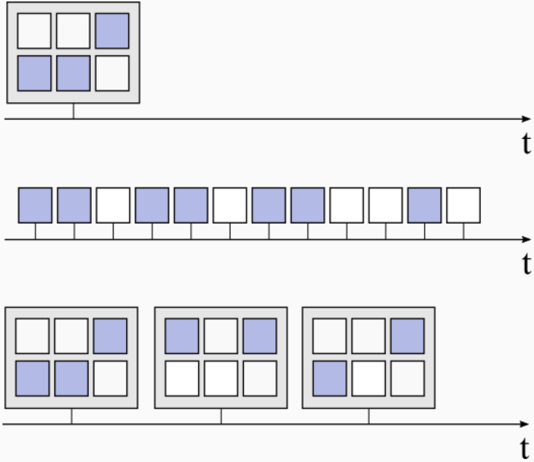


Figure 1: Three types of binary classification datasets — static, incremental, batch. Adapted from [HPC12]

Concept drift

Traditional ML assumption: The dataset is generated by a single, static, hidden function.

Streaming data challenge: The above assumption could be invalid. $f_t(\cdot)$ may be different from $f_{t+1}(\cdot)$. This potential violation is known as *concept drift*.

Where drift occurs

Standard classification problem

Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t\}$ and $\mathbf{y} = \{c_1, c_2, \cdots, c_t\}$ where $\mathbf{x}_i \in \mathbb{R}^d$

$c_i = f(\mathbf{x}_i)$

$\hat{c}_i = \hat{f}(\mathbf{x}_i)$

Concept drift occurs when $f(\mathbf{x}_i)$ changes over time (i.e. $f_t \neq f_{t+1}$)

The standard approach when we know $p(c_i)$ and $p(\mathbf{x}_i|c_i)$ is

$$p(c_i|\mathbf{x}_i) = \frac{p(c_i)p(\mathbf{x}_i|c_i)}{p(\mathbf{x}_i)} \tag{1}$$

Class imbalance

Drift in $p(c_i)$ is related to *class imbalance*. This is bad because

- It hurts the interpretation
- Can be catastrophic in data streams
- Is relevant for static datasets as well
- Can mask concept drift

Drift in Linear regression

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ and
$\mathbf{y} = \{y_1, y_2, \cdots, y_n\}$ where $\mathbf{x}_i \in \mathbb{R}^D, y_i \in \mathbb{R}$
$y_i = \alpha + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \cdots + \beta_n * x_{in} + \epsilon$

$$\hat{\alpha} = \bar{\mathbf{y}} - \hat{\beta}\bar{\mathbf{x}} \tag{2}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})}{(x_i - \bar{\mathbf{x}})^2} \tag{3}$$

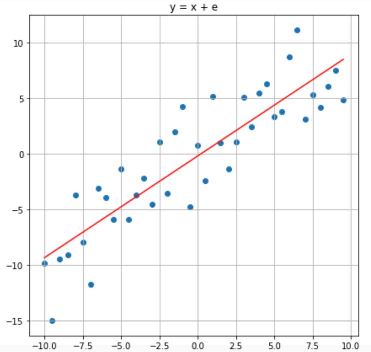


Figure 2: Simple linear regression for D=1

Profiling the training data

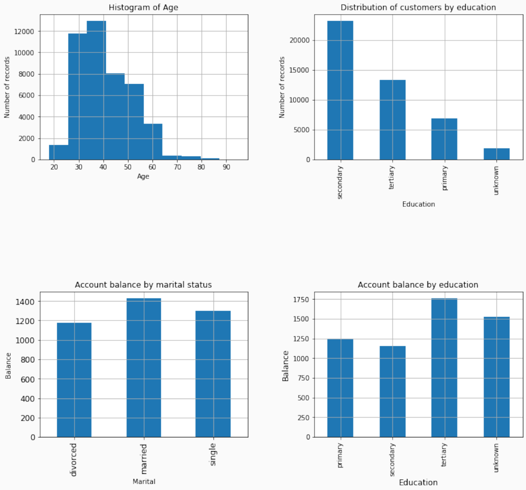
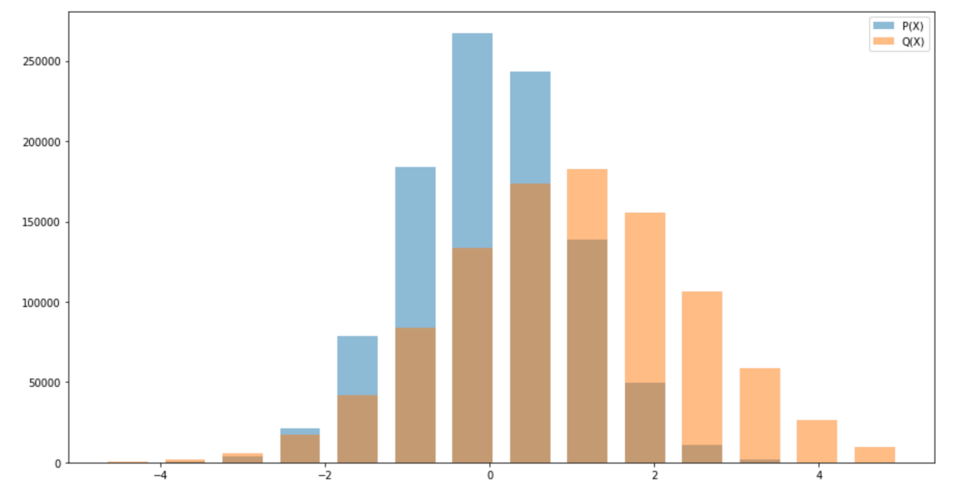


Figure 3: Histograms of four attributes from [MCR14]

Detecting drift



KL Divergence

Kullback-Leibler Divergence

- Measures how one probability distribution is different from another (reference) distribution
- Works well in practice [SG07]

For discrete P and Q defined on the same probability space $\mathcal{X}$ we have:

$$D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{4}$$

KL Divergence

Example[1/2]

Let $P \sim B(n = 2, p = 0.4)$ and $Q \sim U$, $\mathcal{X} \in \{0, 1, 2\}$

x	0	1	2
P(x)	0.358	0.482	0.160
Q(x)	0.333	0.333	0.333

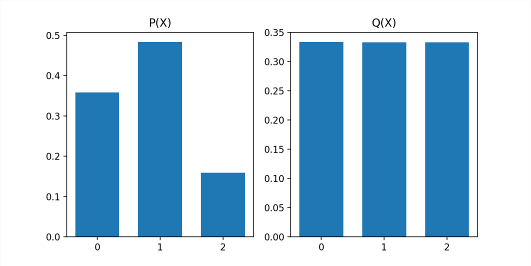


Figure 4: Histograms of $P(X)$ and $Q(X)$

Example[2/2]

x	0	1	2
P(x)	0.358	0.482	0.160
Q(x)	0.333	0.333	0.333

$$
\begin{aligned}
D_{KL}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} &= 0.358 \times \log \frac{0.358}{0.333} + 0.482 \times \log \frac{0.482}{0.333} \\
&+0.358 \times \log \frac{0.160}{0.333} \approx 0.087
\end{aligned} \tag{5}
$$

Speed of drift

Consider a change of the concept generating function from $f(\cdot) \to g(\cdot)$. This change can be

- *abrupt* — $f(\cdot)$ is replaced by $g(\cdot)$ at time step t
- *gradual* — smooth transition from sampling using $f(\cdot)$ to sampling using $g(\cdot)$
- *reoccurring* — either abrupt or gradual, periodical or random

References i

- T. Ryan Hoens, Robi Polikar, and Nitesh V. Chawla, *Learning from streaming data with concept drift and imbalance: an overview*, Progress in Artificial Intelligence **1** (2012), no. 1, 89–101.
- Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin, *The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects*, Frontiers in psychology **4** (2013), 504–504.
- Sérgio Moro, Paulo Cortez, and Paulo Rita, *A data-driven approach to predict the success of bank telemarketing*, Decision Support Systems **62** (2014).

- M. D. Muhlbaier, A. Topalis, and R. Polikar, *Learn^{++} .nc: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes*, IEEE Transactions on Neural Networks **20** (2009), no. 1, 152–168.
- Raquel Sebastião and João Gama, *Change detection in learning histograms from data streams*, 12 2007, pp. 112–123.