Practical tricks

Machine Learning and Data Mining

Maxim Borisyak
National Research University Higher School of Economics (HSE)

October 10, 2017

Outline

Here, we consider practical problems that are not quite aligned with theory:

- imbalanced datasets;
- differences in training and application domains;
- one-class classification.

Imbalanced datasets

Imbalanced datasets

Settings:

- classification problem: $\mathcal{C}^+$ against $\mathcal{C}^-$;
- often in practice $P(\mathcal{C}^+) \ll P(\mathcal{C}^-)$.

This poses several problems:

- mini-batch learning procedures degradate;
  - extreamely slow learning;
- imprecise results.

Degradation of mini-batch learning

Probability of a example from $\mathcal{C}^+$ being selected into a mini-batch is low:

- $\Rightarrow$ increased $\mathbb{D}[\nabla\mathcal{L}]$;
- $\Rightarrow$ low learning rate;
- $\Rightarrow$ slow learning.

Don't train on 50-50 for imbalanced datasets!

$$P(\mathcal{C}^+ \mid X) = \frac{P(X \mid \mathcal{C}^-)P(\mathcal{C}^-)}{P(X \mid \mathcal{C}^-)P(\mathcal{C}^-) + P(X \mid \mathcal{C}^+)P(\mathcal{C}^+)}$$