distribution shift

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

dataset shift, domain shift

Abstract

Distribution shift is a mismatch between the probability distribution underlying the training set and test set of a machine learning (ML) method and the probability distribution of the data points that the learned hypothesis faces after deployment. The average loss on the test set estimates the risk under the training probability distribution only, so a small test error is no evidence for a small loss on data points outside the test set. Factoring the joint probability distribution of features and labels separates the named special cases of covariate shift and label shift.

Definition

The Finnish Meteorological Institute (FMI) station Helsinki Kaisaniemi records the minimum and the maximum air temperature of each day. Every summer day is a data point: its feature is the day's minimum temperature, its label the day's maximum. A linear hypothesis $\widehat{h}(x) = 0.481\,x + 14.5$, learned from 62 days of June–August 2026 by minimizing the average squared error loss, reaches a training error of 5.0 and a test error of 6.5 on a test set of the 30 remaining summer days. Deployed at the northernmost FMI station, Utsjoki Nuorgam (70.1° N), on the 90 days of December 2025–February 2026, the same hypothesis incurs an average error of 242.1, which is 37 times its test error (Fig. 1). Every winter morning at Nuorgam is colder than each summer morning in the training set, and the extrapolation overshoots: the mean winter day at Nuorgam has a minimum of $-19.8\,^{\circ}\mathrm{C}$ and a maximum of $-9.6\,^{\circ}\mathrm{C}$, while $\widehat{h}$ assigns such a morning a maximum of $+5.0\,^{\circ}\mathrm{C}$.

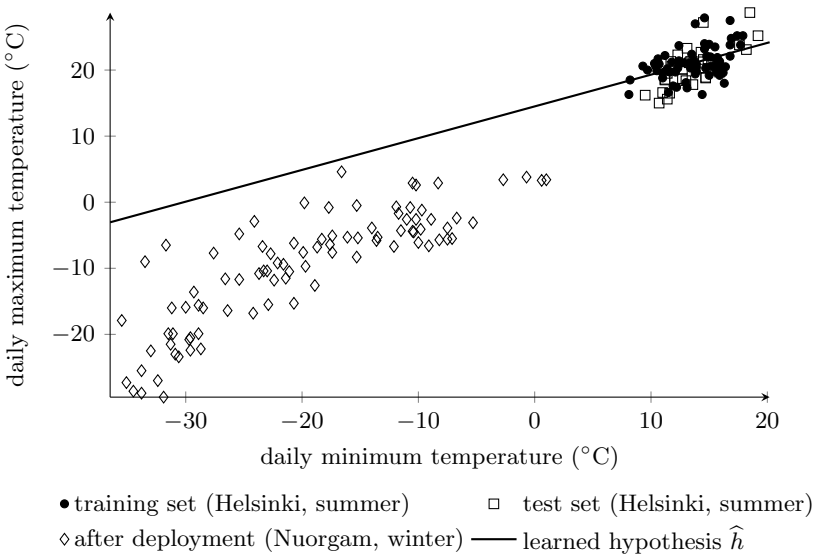

Figure 1: Distribution shift between two FMI stations. Each marker is one day, its feature the daily minimum and its label the daily maximum temperature; the line is the hypothesis $\widehat{h}$ learned from the training set. The features faced after deployment (open diamonds) lie outside the range covered by the training set (filled circles), and $\widehat{h}$ overestimates every winter maximum, by 14.6 °C on average. Data generated by `pythondemos/distshift.py`

Distribution shift is a mismatch between the probability distribution p underlying the training set and test set of a machine learning (ML) method and the probability distribution p' of the data points the learned hypothesis faces after deployment (Quiñonero-Candela et al., 2009). The average loss on the test set estimates the risk under p, since the independent and identically distributed assumption (i.i.d. assumption) covers draws from p, not from p'. A small test error is therefore no evidence about the loss incurred under $p' \neq p$: the densities in Fig. 2 show how a threshold placed for p fails under p', and the two stations of Fig. 1 realize the failure with measured temperatures.

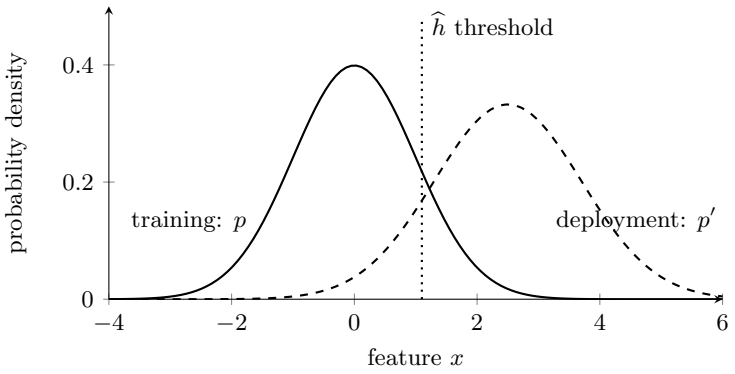

Figure 2: Distribution shift for a single feature. The solid density is the probability distribution p that generated the training set, here a standard Gaussian; the dashed one is the probability distribution p' at deployment, a Gaussian with mean 2.5 and standard deviation 1.2. The dotted threshold of the learned hypothesis $\widehat{h}$ at $x = 1.1$ leaves 14% of the mass of p beyond it, against 88% of the mass of p', so the risk under p' exceeds the risk under p that the test error estimates

The named special cases of distribution shift correspond to the two factorizations of the joint probability distribution of a data point $(\mathbf{x}, y)$ (Moreno-Torres et al., 2012). Under the factorization $p(\mathbf{x}, y) = p(y \mid \mathbf{x})\, p(\mathbf{x})$, covariate shift changes only the marginal $p(\mathbf{x})$ of the covariates (the features) and leaves the conditional $p(y \mid \mathbf{x})$ untouched, as when a second weather station sits in a colder climate but cold mornings precede rain in the same way; concept drift changes that conditional itself over time, as when a warming climate alters which morning temperatures precede rain. Under the reverse factorization $p(\mathbf{x}, y) = p(\mathbf{x} \mid y)\, p(y)$, label shift changes only the frequency $p(y)$ of the label values and leaves the class-conditional $p(\mathbf{x} \mid y)$ untouched. The two-station example is a pure case of none of the three: the marginal of the feature moves, since winter mornings are colder, and the conditional moves as well, since a line fitted to the winter days has slope 0.77 against the summer slope 0.48.

Distribution shift is detected by comparing summary statistics of the features between the training set and the data points arriving after deployment, or by monitoring the incurred loss whenever labels become available (Quiñonero-Candela et al., 2009). One response is sample weighting: each data point of the training set is weighted by the ratio $p'(\mathbf{x})/p(\mathbf{x})$, which corrects covariate shift provided that the conditional $p(y \mid \mathbf{x})$ is unchanged and that every feature vector occurring under p' also occurs under p, so that the ratio is defined (Quiñonero-Candela et al., 2009). Both conditions fail in the two-station example: no reweighting of summer days produces a winter morning at $-20\,^{\circ}\mathrm{C}$. Responses that remain applicable are transfer learning and fine-tuning on data points from p', and online learning, which updates the hypothesis as data points from the

drifting probability distribution arrive (Quiñonero-Candela et al., 2009).

Cross References

generalization, i.i.d. assumption, probability distribution, risk, training set, test set, transfer learning, sample weighting, online learning

References

1. Moreno-Torres et al. (2012). A Unifying View on Dataset Shift in Classification. Pattern Recognition, 45(1), pp. 521–530.

2. Quiñonero-Candela et al. (2009). Dataset Shift in Machine Learning. MIT Press.