

data augmentation

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

Editors: Konstantina Olioumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Data augmentation enlarges a training set with synthetic data points obtained by perturbing or transforming the measured ones. It is one of the three routes that regularization distinguishes, beside shrinking the hypothesis space and adding a penalty term. A perturbation is admissible when it is known to leave the label alone, either because the feature space carries a symmetry, such as the rotation of an image, or because the measurement is only accurate to within a stated tolerance. Enlarging the training set this way removes hypothesis maps that the training error alone cannot rule out, since a copy displaced along the feature axis misses its label whenever the hypothesis is steep there. For the squared error loss, perturbing the features with noise is equivalent to adding a penalty term that grows with the noise variance, which is why the three routes often coincide.

Definition

A weather station at Krems an der Donau measured a maximum temperature of 3.5°C on 21 January 2024, 20.7°C on 15 April and 30.2°C on 19 July. On the day after each of them it measured 0.0 , 13.6 and 27.0°C . Take these three days as a training set for predicting tomorrow's maximum temperature from today's, and let the hypothesis space be the polynomials of degree five (see polynomial regression).

A polynomial of degree five carries six model parameters and the training set holds three data points, so empirical risk minimization (ERM) has infinitely many solutions. Three independent directions change the polynomial without moving it at the three measured temperatures, so every solution has a training error of zero and the training error cannot choose among them. Between the measured days the solutions take every value: at a today-maximum of 12°C the smoothest of them predicts 6.8°C for tomorrow, while the solution drawn in Fig. 1 predicts 21.0°C . The two differ by up to 15°C between the measured days. This is overfitting: the hypothesis space is too large for three data points.

What rules the wild solutions out is something the thermometer already says. It reports the air temperature only to within its accuracy, taken here to be 0.5°C . Both the feature and the label of one of these data points are readings of that instrument, so a day whose two numbers are moved anywhere

inside that tolerance is just as consistent with what was measured. Replacing each measured day by copies drawn that way gives an augmented training set (the open squares in Fig. 1). The perturbation is justified by the instrument, not invented: it asserts only that the labels of two days whose readings agree to within the sensor accuracy should agree too, which is the smoothness assumption read at the scale of the measurement error.

The copies carry many distinct features, so the feature matrix of the augmented training set has full column rank and ERM on it has a unique solution. A polynomial that is steep where the copies lie pays for that steepness, because a copy displaced along the feature axis then misses its label. On the 309 days of 2024 that were held out and whose today-maximum lies between 3.5 and 30.2°C, the solution drawn in Fig. 1 incurs an average squared error loss of 71.8 squared degrees. The solution on the augmented training set incurs 32.6, against 30.9 for the straight line through the three measured days.

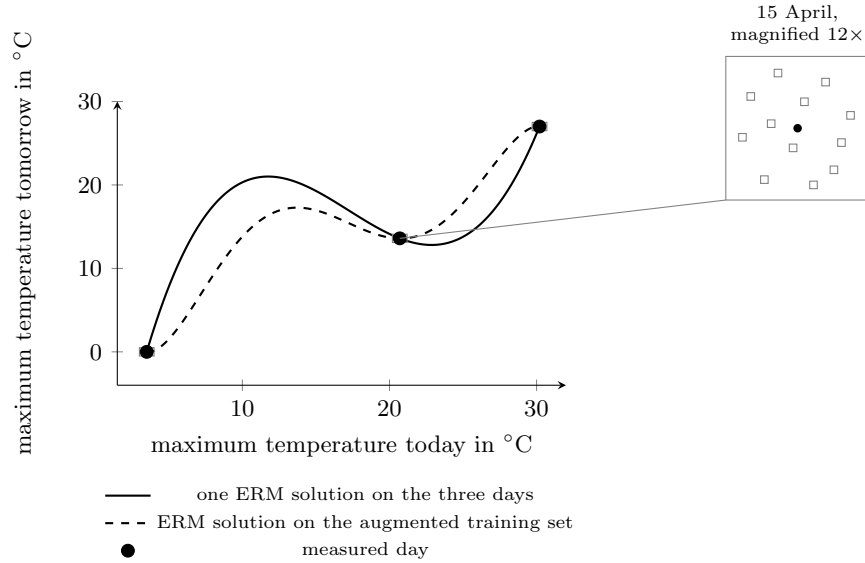


Figure 1: Three measured days at Krems an der Donau, the feature of each being the maximum temperature of that day and the label the maximum temperature of the next. The solid curve is one of the infinitely many polynomials of degree five with a training error of zero on them, and the dashed curve is the unique ERM solution on the augmented training set. The gray box around each measured day is the 0.5°C accuracy of the thermometer in both coordinates. The box at the upper right magnifies that of 15 April twelve times and shows twelve of its copies as open squares, the measured day itself as a filled circle. Data generated by `pythondemos/dataaug.py`

In general, data augmentation methods add synthetic data points to an existing set of data points. These synthetic data points are obtained by pertur-

bations (e.g., adding noise to physical measurements) or transformations (e.g., rotations of images) of the original data points. Enlarging the training set with perturbed copies is one of the three routes that regularization distinguishes, beside shrinking the hypothesis space and adding a penalty term. The three are often equivalent: for the squared error loss, perturbing the features with noise of variance σ^2 is equivalent to adding a penalty term proportional to σ^2 (Bishop, 1995), and replacing every data point by a probability distribution centered at it is the vicinal form of ERM (Chapelle et al., 2001).

When the features are perturbed, the operation is chosen so that the synthetic data points keep the same label. An operator that maps the feature space $\mathcal{X}$ into itself and is known to leave the label alone is called a symmetry of the learning task. For example, a rotated cat image still shows a cat, while its feature vector (obtained by stacking pixel color intensities) differs from that of the original image (see Fig. 2).

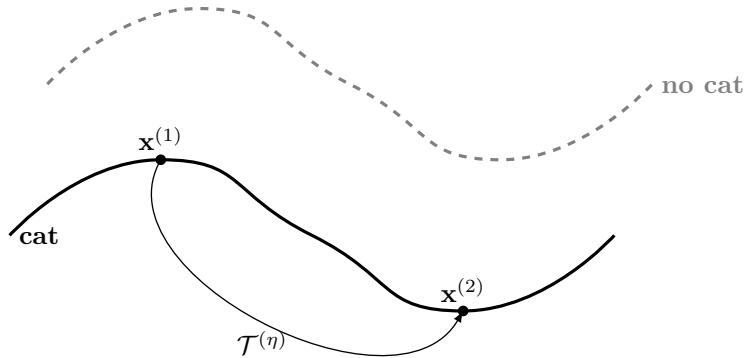


Figure 2: Data augmentation exploits symmetries of data points in the feature space $\mathcal{X}$. The solid curve is the set of feature vectors labelled as a cat, the dashed one a set labelled otherwise. A symmetry is represented by an operator $\mathcal{T}^{(\eta)} : \mathcal{X} \rightarrow \mathcal{X}$, parameterized by some number $\eta \in \mathbb{R}$. For example, $\mathcal{T}^{(\eta)}$ might represent the effect of rotating a cat image by η degrees. A data point with feature vector $\mathbf{x}^{(2)} = \mathcal{T}^{(\eta)}(\mathbf{x}^{(1)})$ must have the same label $y^{(2)} = y^{(1)}$ as a data point with feature vector $\mathbf{x}^{(1)}$

Augmentation can instead perturb the label of a data point. Replacing the labels of training data points with randomly softened or flipped values (label smoothing) is also a form of regularization: it discourages the machine learning (ML) method from fitting individual labels too closely and improves robustness to label noise. For example, logistic regression on a linearly separable training set drives the model parameter vector $\mathbf{w}$ to grow without bound under hard labels: the logistic loss keeps decreasing as the predictions are pushed toward 0 and 1. Smoothing the labels caps the optimal predictions away from 0 and 1, so $\mathbf{w}$ stays finite.

Cross References

regularization, overfitting, ERM, training set, penalty term, hypothesis space, data point, label, feature vector, feature space, smoothness assumption, polynomial regression, ridge regression

References

1. Bishop (1995). Training with Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1), pp. 108–116.
2. Chapelle et al. (2001). Vicinal Risk Minimization. In: *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, pp. 416–422. MIT Press.