Learning by Marginalizing Corrupted Features

Laurens van der Maaten LVDMAATEN@GMAIL.COM
Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

Minmin Chen MC15@CEC.WUSTL.EDU
Stephen Tyree SWTYREE@WUSTL.EDU
Kilian Weinberger KILIAN@WUSTL.EDU
Washington University in St. Louis, St. Louis, MO 63130, USA

Keywords: Regularization, supervised learning, data corruption.

Overfitting is a key problem in machine learning. It is generally combatted using *regularization* or *Bayesian* techniques that employ priors favoring "simple" models over "complex" models. We propose a new approach to counter overfitting, called *marginalized corrupted features* (van der Maaten et al., 2013, MCF). Instead of perturbing models, which can be counter-intuitive, MCF regularizes by *perturbing the data*. We may know that certain *corruptions* of data instances do not affect their label. For example, deleting a few words in a document rarely changes its topic. MCF uses this knowledge to generate additional data that looks like real data: it corrupts the *finite* training set with a corrupting distribution to construct an *infinite* corrupted training set on which the model is trained.

The corrupting distribution, which specifies how training observations $\mathbf{x}$ are transformed into corrupted versions $\tilde{\mathbf{x}}$, is assumed to factorize over dimensions d:

$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \prod_{d=1}^{D} P_E(\tilde{x}_d|x_d;\eta_d).$$

Herein, η_d represents hyperparameters of the corrupting distribution. Corrupting distributions of interest, P_E, include: (1) "blankout" noise in which features are randomly set to zero, (2) Gaussian noise, and (3) Poisson noise in which features are used as rates.

Assume we have a training set $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ and a loss function $L(\mathbf{x}, y; \Theta)$, with model parameters Θ. A simple approach to use the corrupting distribution is by corrupting each training sample M times, and training on the resulting NM corrupted instances. Such an approach is effective (Vincent et al., 2008), but it lacks elegance and is computationally expensive. MCF addresses these issues by considering the limiting case $M \to \infty$, in which we obtain the *expected loss* under the corrupting distribution:

$$\mathcal{L}(\mathcal{D};\Theta) = \sum_{n=1}^{N} \mathbb{E}[L(\tilde{\mathbf{x}}_n, y_n; \Theta)]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)}.$$

For linear models, this expectation can be computed analytically for many loss functions and corrupting distributions. For a linear model with weights $\mathbf{w}$, the expected value of the *quadratic loss* under the corrupting distribution $p(\tilde{\mathbf{x}}|\mathbf{x})$ is (Chen et al., 2012):

$$\begin{aligned}\mathcal{L}(\mathcal{D};\mathbf{w}) &= \sum_{n=1}^{N} \mathbb{E}\left[\left(\mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n - y_n\right)^2\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)} \\ &= \mathbf{w}^{\mathrm{T}}\mathbf{H}\mathbf{w} - 2\left(\sum_{n=1}^{N} y_n \mathbb{E}\left[\tilde{\mathbf{x}}_n\right]\right)^{\mathrm{T}}\mathbf{w} + N,\end{aligned}$$

where the hat matrix $\mathbf{H} = \sum_{n=1}^{N} \mathbb{E}[\tilde{\mathbf{x}}_n]\mathbb{E}[\tilde{\mathbf{x}}_n]^{\mathrm{T}} + V[\tilde{\mathbf{x}}_n]$, and $V[x]$ is the variance of x. Hence, to minimize the expected quadratic loss, we only need to compute the mean and variance of the corrupting distribution. This is efficient for a wide range of corruption models.

For *logistic loss*, we derive a closed-form upper bound:

$$\begin{aligned}\mathcal{L}(\mathcal{D};\mathbf{w}) &= \sum_{n=1}^{N} \mathbb{E}\left[\log\left(1 + \exp\left(-y_n \mathbf{w}^{\mathrm{T}}\tilde{\mathbf{x}}_n\right)\right)\right]_{p(\tilde{\mathbf{x}}_n|\mathbf{x}_n)} \\ &\leq \sum_{n=1}^{N} \log\left(1 + \prod_{d=1}^{D} \mathbb{E}\left[\exp\left(-y_n w_d \tilde{x}_{nd}\right)\right]_{p(\tilde{x}_{nd}|x_{nd})}\right).\end{aligned}$$

Herein, we recognize a product of moment-generating functions that can be computed efficiently for corrupting distributions in the natural exponential family.

We show the merits of learning with MCF for various models and corrupting distributions. In particular, MCF achieves substantial performance improvements in document classification and domain adaptation.

References

Chen, M., Xu, Z., Weinberger, K., & Sha, F. (2012). Marginalized denoising autoencoders for domain adaptation. *Proceedings of the International Conference on Machine Learning* (pp. 767–774).

van der Maaten, L., Chen, M., Tyree, S., & Weinberger, K. (2013). Learning by marginalizing corrupted features. *Proceedings of the International Conference on Machine Learning, JMLR W&CP*, *28*, 410–418.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the International Conference on Machine Learning* (pp. 1096–1103).