Measuring the effect of nuisance variables on classifiers

Alhussein Fawzi
alhussein.fawzi@epfl.ch
Pascal Frossard
pascal.frossard@epfl.ch

Signal Processing Laboratory (LTS4)
Ecole Polytechnique Fédérale de
Lausanne (EPFL)
Lausanne, Switzerland

1 Proof of Theorem 1

Theorem 1. Let $t > 0$, and $\delta \in (0, 1)$. We have $|\hat{\rho}_T - \rho_T| \leq t$ with probability exceeding $1 - \delta$ as long as

$$M \geq \frac{\ln(2/\delta)}{2t^2}.$$ 

Moreover, when the prior distributions are data-independent (i.e., $p_T(\theta|x) = p_T(\theta)$), the condition in Eq. (1) becomes

$$NM \geq \frac{\ln(2/\delta)}{2t^2}.$$ 

Proof. Our main ingredient for proving this result is Hoeffding’s inequality. We recall this inequality as follows:

Theorem 2 (Hoeffding’s inequality). Let $(X_i, i \geq 1)$ be a sequence of independent random variables such that $0 \leq X_i \leq 1$. If $\bar{X}_n = \frac{1}{n} (X_1 + \cdots + X_n)$, then for all $t > 0$

$$\mathbb{P}(\{|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq t\}) \leq 2 \exp\left(-\frac{2t^2}{n}\right).$$

Case (a). We start our proof by considering the case where the prior distribution does not depend on the image: $p_T(\theta|x) = p_T(\theta)$, to establish the result in Eq. (2). We have:

$$\rho_T = \int_x \int_\theta p_{\text{cl}}(\ell(x)|x, \theta)p_T(\theta)p_d(x)d\theta dx,$$

$$\hat{\rho}_T = \frac{1}{M} \sum_{j=1}^M \frac{1}{N} \sum_{i=1}^N p_{\text{cl}}(\ell(x_j)|x_j, \theta_i) := \frac{1}{M} \frac{1}{N} \sum_{j=1}^M \sum_{i=1}^N Z_{j,i}.$$ 

The random variables $\theta_i$ and $x_j$ are independent, hence $\{Z_{j,i}\}_{(j,i)}$ are pairwise independent. Note moreover that $Z_{j,i} \in [0, 1]$, and that $\mathbb{E}(Z_{j,i}) = \rho_T$ for any $j, i$. Hence, by applying Hoeffding’s inequality, we obtain

$$\mathbb{P}(|\hat{\rho}_T - \rho_T| \geq t) \leq 2 \exp\left(-2NMt^2\right).$$

© 2016. The copyright of this document resides with its authors. It may be distributed unchanged freely in print or electronic forms.
Setting $\delta = 2 \exp(-2NMt^2)$, we obtain the desired result in Eq. (2).

**Case (b).** We now consider the general case where the the prior distribution $p_T(\theta|x)$ depends on the image, and our goal is to establish the result in Eq. (1). We have:

$$\rho_T = \int_x \int_{\theta} p_c(\ell(x)|x, \theta)p_T(\theta|x)p_d(x)d\theta dx,$$

$$\hat{\rho}_T = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N} \sum_{i=1}^{N} p_c(\ell(x_j)|x_j, \theta_i) = \frac{1}{M} \frac{1}{N} \sum_{j=1}^{M} \sum_{i=1}^{N} Z_{j,i}.$$

In this case, the random variables $Z_{j,i}$ and $Z_{j,i'}$ might be dependent (for $i \neq i'$), as $\theta_i$ and $\theta_{i'}$ are only conditionally independent. We therefore introduce the random variable

$$W_j = \frac{1}{N} \sum_{i=1}^{N} Z_{j,i},$$

and note that $\{W_j\}_j$ are pairwise independent, as the random variables $\{x_j\}$ are chosen independently. Note moreover that $\mathbb{E}(W_j) = \mathbb{E}(Z_{j,i}) = \rho_T$, and that $W_j \in [0, 1]$. We apply Hoeffding’s inequality for $W_j$ and obtain

$$\mathbb{P}(|\hat{\rho}_T - \rho_T| \geq t) \leq 2 \exp(-2Mt^2).$$

By setting $\delta = 2 \exp(-2Mt^2)$, we obtain the desired result in Eq. (1). □

## 2 Additional experimental description and illustrations

### 2.1 MNIST handwritten digits

In this experiment, the nuisance set $\mathcal{T}$ is the set of affine transformations. We parametrize each element $\mathcal{T}$ with a vector $\theta \in \mathbb{R}^6$. We impose a Gaussian prior $p_T(\cdot|x) = \mathcal{N}(1, \Sigma)$, where $1$ denotes the identity transformation, and $\Sigma$ denotes the covariance matrix. We set the covariance matrix in order to penalize large changes in the *appearance* of the image. The covariance therefore naturally depends on the image $x$, since, for example, the appearance of a circular image is not altered under the action of rotations. To define the notion of *appearance change*, we follow a similar approach to that of [1, 2, 3]. We quantify the change in appearance between two elements $\theta_0$ and $\theta_1$ in $\mathcal{T}$ using the geodesic distance on the manifold of transformed samples $\{T_{\theta}x : \theta \in \mathcal{T}\}$. This distance can be written

$$d(\theta_0, \theta_1) = \inf_{\gamma} \int_0^1 \sqrt{\gamma(t)^T G_{\gamma(t)} \gamma(t)} dt,$$

where the infimum is taken over all $C^1$ curves $\gamma$ that satisfy $\gamma(0) = \theta_0$ and $\gamma(1) = \theta(1)$, and $G$ denotes a Riemannian metric on the manifold $\mathcal{T}$ [4]. When $\theta_1$ is in the neighborhood of $\theta_0$, we can approximate the matrix $G_{\gamma(t)}$ (for any $t$) by $G_{\theta_0}$, provided $G_{\gamma(t)}$ is slowly varying with $\gamma(t)$. By assuming a constant $G_{\gamma(t)} = G_{\theta_0} = G$, the distance in Eq. (3) can be computed in closed-form. It is easy to see that when $G_{\gamma(t)}$ is constant, we have

$$d(\theta_0, \theta_1) = \sqrt{(\theta_1 - \theta_0)^T G(\theta_1 - \theta_0)}.$$
We naturally set the prior distribution on $\mathcal{T}$ in order to penalize large variations in the appearance of the image, by defining

$$p_{\mathcal{T}}(\theta|x) \propto \exp\left(-\alpha d(1, \theta)^2\right) = \exp\left(-(\theta - 1)^T \Sigma^{-1} (\theta - 1)\right),$$

with $\Sigma^{-1} = \alpha G$, and $\alpha$ is a parameter controlling the “magnitude” of the transformation. In that sense, our prior distribution hence penalizes changes in appearance of the image, and favors nuisance regions that do not significantly distort the data.

We show in Fig. 1 transformed versions of arbitrary MNIST images with nuisance samples drawn from the prior $p_{\mathcal{T}}(\theta|x)$, for $\alpha = 100, 50, 10$.

![Figure 1: Original images are shown in row 1. Samples drawn from prior distribution with $\alpha = 100$ [row 2, mild transformations], $\alpha = 50$ [row 3, medium transformations], and $\alpha = 10$ [row 4, severe transformations].](image)

2.2 Natural images & face recognition

In Fig. 2, we show samples from the prior distribution $p_{\mathcal{T}}(\theta)$ (the prior is independent of $x$ here), when $\mathcal{T}$ is the set of piecewise affine transformations, for randomly taken images in the ILSVRC 2012 validation set.

![Figure 2: Transformed versions of images taken from the ILSVRC 2012 validation dataset.](image)
References

