Unsupervised classifier risk derivation

Christophe Cerisara

Unsupervised classifier risk derivation

Christophe Cerisara
Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

February 18, 2019

1 Introduction

The following derivation is a supplemental material for an article that has been submitted to a conference. This derivation is quite simple, but it gives a closed-form expression for an unsupervised binary classifier risk approximation that can be included within any deep neural network model and be used to train this model in an unsupervised way with standard backpropagation and stochastic gradient descent.

2 Derivation of the unsupervised risk

Let be given a binary linear classifier with parameters θ that computes a scalar score $f(x) = \sum_{i=1}^{n} \theta_i x_i \in \mathbb{R}$ for observation $x \in \mathbb{R}^n$. The classifier outputs class $\hat{y} = 0$ iff $f(x) < 0$, and $\hat{y} = 1$ iff $f(x) >= 0$. The true/gold class label is noted $y \in \{0, 1\}$. The risk of this classifier with a hinge loss is [1]:

$$\begin{aligned} R(\theta) &= E_{p(x,y)}\left[(1 - f(x) \cdot (2y-1))_+\right] \\ &= P(y=0) \int p(f(x) = \alpha | y = 0)(1+\alpha)_+ d\alpha + \\ &\quad P(y=1) \int p(f(x) = \alpha | y = 1)(1-\alpha)_+ d\alpha \end{aligned} \tag{1}$$

We assume the conditional distributions follow a normal distribution:

$$p(f(x)|y=0) \sim N(\mu_0, \sigma_0)$$

$$p(f(x)|y=1) \sim N(\mu_1, \sigma_1)$$

where $N(\mu, \sigma)$ is the standard normal distribution with mean μ and variance σ^2. We can then rewrite an approximation of this risk under the previous assumption and the additional assumption that the class marginals $P(y)$ are known:

$$R = P(y=0) \int N(\alpha; \mu_0, \sigma_0)(1+\alpha)_+ d\alpha + P(y=1) \int N(\alpha; \mu_1, \sigma_1)(1-\alpha)_+ d\alpha$$

Removing the non-linearity:

$$R = P(y=0)\int_{-1}^{+\infty} N(\alpha;\mu_0,\sigma_0)(1+\alpha)d\alpha + P(y=1)\int_{-\infty}^{1} N(\alpha;\mu_1,\sigma_1)(1-\alpha)d\alpha$$

Distributing:

$$\begin{aligned} R &= P(y=0)\int_{-1}^{+\infty} N(\alpha;\mu_0,\sigma_0)d\alpha + P(y=0)\int_{-1}^{+\infty} \alpha N(\alpha;\mu_0,\sigma_0)d\alpha + \\ & P(y=1)\int_{-\infty}^{1} N(\alpha;\mu_1,\sigma_1)d\alpha - P(y=1)\int_{-\infty}^{1} \alpha N(\alpha;\mu_1,\sigma_1)d\alpha \end{aligned}$$

We know that the cumulative distribution function of a (μ,σ) normal is:

$$F(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

So the integral of a Gaussian is:

$$\int_a^b N(x;\mu,\sigma)dx = F(b) - F(a) = \frac{1}{2}\left(\text{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right)$$

We know that $\text{erf}(-\infty) = -1$, $\text{erf}(0) = 0$ and $\text{erf}(+\infty) = 1$, so

$$\begin{aligned} R &= \frac{P(y=0)}{2}\left(1 - \text{erf}\left(\frac{-1-\mu_0}{\sigma_0\sqrt{2}}\right)\right) + P(y=0)\int_{-1}^{+\infty} \alpha N(\alpha;\mu_0,\sigma_0)d\alpha + \\ & \frac{P(y=1)}{2}\left(1 + \text{erf}\left(\frac{1-\mu_1}{\sigma_1\sqrt{2}}\right)\right) - P(y=1)\int_{-\infty}^{1} \alpha N(\alpha;\mu_1,\sigma_1)d\alpha \end{aligned}$$

Integration by parts give:

$$\int_a^b uv'dx = [uv]_a^b - \int_a^b u'vdx$$

So with $u = x$ and $v' = N(x;\mu,\sigma)$

$$\int_a^b xN(x;\mu,\sigma)dx = bF(b) - aF(a) - \int_a^b F(x)dx$$

We also know that

$$\int \text{erf}(x)dx = x\text{erf}(x) + \frac{e^{-x^2}}{\sqrt{\pi}} + C$$

So

$$\int_a^b F(x)dx = \frac{b-a}{2} + \frac{1}{2}\int_a^b \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)dx$$

We use integration by substitution:

$$\int_a^b \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)dx = \sigma\sqrt{2}\int_a^b \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\frac{dx}{\sigma\sqrt{2}} = \sigma\sqrt{2}\int_{u(a)}^{u(b)} \operatorname{erf}(u)du$$

with $u(x) = \frac{x-\mu}{\sigma\sqrt{2}}$

So

$$\int_{u(a)}^{u(b)} \operatorname{erf}(u)du = u(b)\operatorname{erf}(u(b)) - u(a)\operatorname{erf}(u(a)) + \frac{1}{\sqrt{\pi}}(e^{-u(b)^2} - e^{-u(a)^2})$$

$$\int_{u(a)}^{u(b)} \operatorname{erf}(u)du = \frac{b-\mu}{\sigma\sqrt{2}}\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \frac{a-\mu}{\sigma\sqrt{2}}\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) + \frac{1}{\sqrt{\pi}}(e^{-\frac{(b-\mu)^2}{2\sigma^2}} - e^{-\frac{(a-\mu)^2}{2\sigma^2}})$$

$$\int_a^b \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)dx = (b-\mu)\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - (a-\mu)\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) + \frac{\sigma\sqrt{2}}{\sqrt{\pi}}(e^{-\frac{(b-\mu)^2}{2\sigma^2}} - e^{-\frac{(a-\mu)^2}{2\sigma^2}})$$

$$\int_a^b F(x)dx = \frac{b-a}{2} + \frac{b-\mu}{2}\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \frac{a-\mu}{2}\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) + \frac{\sigma}{\sqrt{2\pi}}(e^{-\frac{(b-\mu)^2}{2\sigma^2}} - e^{-\frac{(a-\mu)^2}{2\sigma^2}})$$

Plugging into the former equation:

$$\int_a^b xN(x;\mu,\sigma)dx = bF(b) - aF(a) - \frac{b-a}{2} - \frac{b-\mu}{2}\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) + \frac{a-\mu}{2}\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) - \frac{\sigma}{\sqrt{2\pi}}(e^{-\frac{(b-\mu)^2}{2\sigma^2}} - e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned}\int_a^b xN(x;\mu,\sigma)dx &= \frac{b}{2} + \frac{b}{2}\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \frac{a}{2} - \frac{a}{2}\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) + \\ &\quad -\frac{b}{2} + \frac{a}{2} - \frac{b}{2}\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) + \frac{\mu}{2}\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) + \\ &\quad \frac{a}{2}\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) - \frac{\mu}{2}\operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) - \frac{\sigma}{\sqrt{2\pi}}(e^{-\frac{(b-\mu)^2}{2\sigma^2}} - e^{-\frac{(a-\mu)^2}{2\sigma^2}})\end{aligned}$$

Simplifying

$$\begin{aligned}\int_a^b xN(x;\mu,\sigma)dx &= \frac{\mu}{2}\left(\operatorname{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right) - \\ &\quad \frac{\sigma}{\sqrt{2\pi}}(e^{-\frac{(b-\mu)^2}{2\sigma^2}} - e^{-\frac{(a-\mu)^2}{2\sigma^2}})\end{aligned}$$

Note: another way to obtain this result is to use the following known formula:

$$\int_a^b xN(x;\mu,\sigma)dx = \mu\int_a^b N(x;\mu,\sigma)dx - \sigma^2\left[N(x;\mu,\sigma)\right]_a^b$$

So

$$\int_a^b xN(x;\mu,\sigma)dx = \frac{\mu}{2}\left(\text{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right) - \sigma^2\left(N(b;\mu,\sigma) - N(a;\mu,\sigma)\right)$$

When $b \to +\infty$:

$$\int_a^{+\infty} xN(x;\mu,\sigma)dx = \frac{\mu}{2}\left(1 - \text{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right) + \sigma^2 N(a;\mu,\sigma)$$

And when $a \to -\infty$:

$$\int_{-\infty}^b xN(x;\mu,\sigma)dx = \frac{\mu}{2}\left(1 + \text{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right)\right) - \sigma^2 N(b;\mu,\sigma)$$

Our risk is:

$$\begin{aligned} R &= \frac{P(y=0)}{2}\left(1 - \text{erf}\left(\frac{-1-\mu_0}{\sigma_0\sqrt{2}}\right)\right) + P(y=0)\int_{-1}^{+\infty} \alpha N(\alpha;\mu_0,\sigma_0)d\alpha + \\ &\quad \frac{P(y=1)}{2}\left(1 + \text{erf}\left(\frac{1-\mu_1}{\sigma_1\sqrt{2}}\right)\right) - P(y=1)\int_{-\infty}^{1} \alpha N(\alpha;\mu_1,\sigma_1)d\alpha \end{aligned}$$

So

$$\begin{aligned} R &= \frac{P(y=0)}{2}\left(1 - \text{erf}\left(\frac{-1-\mu_0}{\sigma_0\sqrt{2}}\right)\right) + -\frac{P(y=0)\mu_0}{2}\left(1 - \text{erf}\left(\frac{-1-\mu_0}{\sigma_0\sqrt{2}}\right)\right) + \\ &\quad P(y=0)\sigma_0^2 N(-1;\mu_0,\sigma_0) + \\ &\quad \frac{P(y=1)}{2}\left(1 + \text{erf}\left(\frac{1-\mu_1}{\sigma_1\sqrt{2}}\right)\right) - \frac{P(y=1)\mu_1}{2}\left(1 + \text{erf}\left(\frac{1-\mu_1}{\sigma_1\sqrt{2}}\right)\right) + \\ &\quad P(y=1)\sigma_1^2 N(1;\mu_1,\sigma_1) \end{aligned}$$

And finally, the risk as a function of the bi-Gaussian parameters is:

$$\begin{aligned} R &= \frac{P(y=0)}{2}(1+\mu_0)\left(1 - \text{erf}\left(\frac{-1-\mu_0}{\sigma_0\sqrt{2}}\right)\right) + P(y=0)\sigma_0^2 N(-1;\mu_0,\sigma_0) + \\ &\quad \frac{P(y=1)}{2}(1-\mu_1)\left(1 + \text{erf}\left(\frac{1-\mu_1}{\sigma_1\sqrt{2}}\right)\right) + P(y=1)\sigma_1^2 N(1;\mu_1,\sigma_1) \end{aligned}$$

We have implemented this risk in pytorch [2] in order to benefit from automatic differentiation to compute the gradient of this risk with respect to the bi-Gaussian parameters.

Then, we propose in the main paper two solutions to connect this gradient to the rest of the neural architecture in order to perform end-to-end unsupervised training:

- The first option exploits two nested SGD processes, respectively for optimizing the bi-Gaussian parameters and the other network parameters;
- The second option approximates the bi-Gaussian parameters estimation with a deterministic function, which allows to directly plug the risk at the output of the deep neural network and set-up a unique backpropagation chain for end-to-end training.

References

[1] Krishnakumar Balasubramanian, Pinar Donmez, and Guy Lebanon. Unsupervised supervised learning II: Margin-based classification without labels. *Journal of Machine Learning Research*, 12:3119–3145, 2011.

[2] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.