Machine learning - Master ICFP 2019-2020

Maximum likelihood and losses for classification

Lénaïc Chizat

February 7, 2020

In this lecture we study how to choose a loss function in supervised machine learning in a principled way.

1 Announcement

Practical session 1 is due February 7, 2020 (today).

2 Maximum Likelihood

Maximum likelihood is a statistical estimation method which can be applied as follows in machine learning:

(i) define a probabilistic model of the data which depends on some parameters (i.e. not fully specified),

(ii) learn/estimate the parameters of the model with the *maximum likelihood principle*,

(iii) use this fully specified model for prediction, generation, or any other task.

2.1 Definition

Let μ be a reference measure on $\mathcal{Z}$ (such as the counting measure on $\mathbb{N}$ or the Lebesgue measure on $\mathbb{R}$). Later, for supervised machine learning problems, we will take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ or $\mathcal{Z} = \mathcal{Y}$.

Definition 1 (Parametric model of distributions) *Let $\Theta \subset \mathbb{R}^p$ be a set of parameters. A probabilistic model $\mathcal{P}$ is a family of probability distributions on $\mathcal{Z}$ which are absolutely continuous with respect to μ and indexed by Θ. We write $\mathcal{P} = \{p_\theta \mathrm{d}\mu \; ; \; \theta \in \Theta\}$.*

Examples: binomial, multinomial, univariate or multivariate Gaussian... The notion of *exponential families* provides a convenient general framework (not studied here).

Definition 2 (Likelihood) *For a data $z \in \mathcal{Z}$, the likelihood is the function $L : \theta \mapsto p_\theta(z)$. Given an i.i.d. training set $(z_i)_{i=1}^n \in \mathcal{Z}^n$, its likelihood is*

$$L(\theta) := \Pi_{i=1}^n p_\theta(z_i).$$

Principle of maximum likelihood: choose the parameters that maximize the likelihood of the training data.

- popularized by Ronald Fisher in the early XXth century,
- shown later to have nice properties (smallest variance over unbiaised estimators),
- maximizing the likelihood L is equivalent to minimizing $-\log L$ because $-\log$ is strictly decreasing. We thus need to solve
$$\min_{\theta\in\Theta} -\log L(\theta) = \min_{\theta\in\Theta} -\sum_{i=1}^n \log(p_\theta(z_i)).$$

2.2 Formulation as a risk minimization

ERM with the log-loss We are in the setting of *density estimation*: given training data, we want to identify the distribution from which it was sampled.

- Consider the *log-loss*: $\ell(z,\theta) = -\log(p_\theta(z))$.
- The corresponding risk is (for random data Z)
$$\mathcal{R}(\theta) = -\mathbb{E}\Big[\log(p_\theta(Z))\Big].$$
- The empirical risk is by definition
$$\hat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^n \log(p_\theta(z_i)) = -\frac{1}{n}\log L(\theta).$$

With this loss, empirical risk minimization coincides with maximum likelihood estimation.

Kullback-Leibler divergence Under the assumption that $Z \sim p_{\theta_0} d\mu$ for some $\theta_0 \in \Theta$, we have

$$\begin{aligned}\mathcal{R}(\theta) - \mathcal{R}(\theta_0) &= -\mathbb{E}\Big[\log(p_\theta(Z))\Big] + \mathbb{E}\Big[\log(p_{\theta_0}(Z))\Big]\\ &= \mathbb{E}\Big[\log\Big(\frac{p_{\theta_0}(Z)}{p_\theta(Z)}\Big)\Big] =: \mathrm{KL}(p_{\theta_0}, p_\theta)\end{aligned}$$

where $\mathrm{KL}(p_\theta, p_{\theta'})$ is the Kullback-Leibler divergence between p_θ and $p_{\theta'}$, also called the entropy of p_θ relative to $p_{\theta'}$. This quantity appears in many fundamental results in probability and statistics. For $\theta, \theta' \in \Theta$, it is defined as

$$\mathrm{KL}(p_\theta, p_{\theta'}) = \int_{\mathcal{Z}} \log\Big(\frac{p_\theta(z)}{p_{\theta'}(z)}\Big) p_\theta(z) d\mu(z).$$

(or $+\infty$ if $p_\theta d\mu$ is not absolutely continuous with respect to $p_{\theta'} d\mu$).

Lemma 1 *For any $\theta, \theta' \in \Theta$, it holds* $\mathrm{KL}(p_\theta, p_{\theta'}) \geq 0$ *and* $\mathrm{KL}(p_\theta, p_{\theta'}) = 0$ *if and only if* $p_\theta = p_{\theta'}$.

Proof Consider the function $\phi(s) = s\log(s) - s + 1$ which is strictly convex and satisfies $\phi(1) = 0$. It holds

$$\int_{\mathcal{Z}} \phi(p_\theta(z)/p_{\theta'}(z))p_{\theta'}(z)d\mu(z) = \int_{\mathcal{Z}} \log\Big(\frac{p_\theta(z)}{p_{\theta'}(z)}\Big)p_\theta(z)d\mu(z) - \int_{\mathcal{Z}} p_\theta(z)d\mu(z) + \int_{\mathcal{Z}} p_{\theta'}(z)d\mu(z) = \mathrm{KL}(p_\theta, p_{\theta'}).$$

By Jensen's inequality

$$\mathrm{KL}(p_\theta, p_{\theta'}) = \int \phi(p_\theta/p_{\theta'})p_{\theta'}d\mu \geq \phi\Big(\int p_\theta d\mu\Big) = \phi(1) = 0.$$

where this inequality is strict unless $p_\theta/p_{\theta'} = 1$ ($p_{\theta'}d\mu$−almost everywhere), i.e. $p_{\theta'} = p_\theta$. ■

As a direct application of this lemma, we have:

Proposition 1 *If $Z \sim p_{\theta_0}d\mu$, then the Bayes risk is $\mathcal{R}^* = \mathcal{R}(\theta_0)$ and the excess risk is*

$$\mathcal{R}(\theta) - \mathcal{R}^* = \mathrm{KL}(p_{\theta_0}, p_\theta).$$

- The proof of the lemma works for any ϕ that is strictly convex and satisfies $\phi(1) = 0$. The resulting quantity is called a ϕ-divergence,
- KL is not a distance as it is not symmetric and does not satisfy a triangular inequality. It is however often helpful to think of it as a "squared distance" over the space of distributions $\mathcal{P}$.

3 Examples of maximum likelihood estimators

Maximum likelihood can be applied to the supervised learning setting, by letting $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$:

$$\mathcal{R}(\theta) = -\mathbb{E}\Big[\log(p_\theta(X, Y))\Big] \qquad\qquad \hat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log(p_\theta(x_i, y_i)).$$

It is also possible to by-pass the need to define a model for the distribution of X and only model the conditional distribution $Y|X$. This gives the conditional log-likelihood risks:

$$\mathcal{R}(\theta) = -\mathbb{E}\Big[\log(p_\theta(Y|X))\Big] \qquad\qquad \hat{\mathcal{R}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log(p_\theta(y_i|x_i)).$$

3.1 Conditional models for linear regression

Linear model with Gaussian noise Consider the following conditional model, with parameters $\theta = (w, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+$

$$Y = X^\top w + Z$$

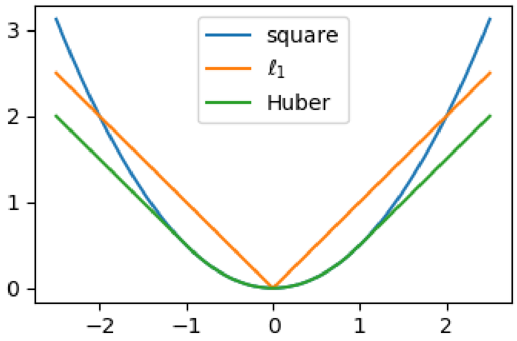

Figure 1: Classical losses used in regression tasks

where $Z \sim \mathcal{N}(0, \sigma^2)$. The likelihood of a training set $(x_i, y_i)_{i=1}^n$ is

$$L(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-(y_i - x_i^\top w)^2/(2\sigma^2)\Big).$$

Thus

$$\hat{\mathcal{R}}(\theta) = -\frac{1}{n} \log L(\theta) = \frac{1}{2}\log(2\pi) + \log\sigma + \frac{1}{2n\sigma^2}\sum_{i=1}^n (y_i - x_i^\top w)^2.$$

Thus the maximum likelihood estimator for w coincides with the *ordinary least squares* estimator.

Exercise: what is the maximum likelihood of σ^2? Solution: $\sigma^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - x_i^\top \hat{w})^2$ (note that this estimator is biased towards 0).

Linear model with Laplace noise Consider the following conditional model, with parameters $\theta = (w, b) \in \mathbb{R}^d \times \mathbb{R}_+$

$$Y = X^\top w + Z$$

where $Z \sim \text{Laplace}(0, b)$. The likelihood of a training set $(x_i, y_i)_{i=1}^n$ is

$$L(\theta) = \prod_{i=1}^n \frac{1}{2b} \exp\Big(-|y_i - x_i^\top w|/b\Big).$$

Thus

$$\hat{\mathcal{R}}(\theta) = -\frac{1}{n}\log L(\theta) = \log(2b) + \frac{1}{bn}\sum_{i=1}^n |y_i - x_i^\top w|.$$

The maximum likelihood estimator here is called the *least absolute error estimator*. It coincides with the empirical risk minimization for linear models and the ℓ_1-loss $\ell(y,z) = |y - z|$. It is called a *robust regression* method because it gives less importance to outliers.

Summing up We thus see that the maximum likelihood principle can serve as a motivation for choosing a loss. See Figure 1 for an illustration. Other commonly used losses "interpolate" between ℓ_1 and ℓ_2: they are locally quadratic around 0 and have the asymptotic growth of the ℓ_1-loss (such as the *Huber loss*).

3.2 Conditional model for linear classification: the logistic loss

Linear log-odds model Let $X \in \mathbb{R}^d$, $Y \in \{0,1\}$ and assume that for a given observation $X = x$, the output $Y|X = x$ follows a Bernoulli law with parameter $\eta(x)$ (i.e. $\eta(x) = P(Y = 1|X = x)$). We make the *linear logit model* assumption: for some $w \in \mathbb{R}^d$, it holds $\forall x \in \mathcal{X}$,

$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = x^\top w$$

where w is again a weight vector (as for linear regression we still can add an offset/intercept in practice). Inverting this function, we get

$$\eta(x) = (1-\eta(x))e^{x^\top w} \Leftrightarrow \eta(x)(1+e^{x^\top w}) = e^{x^\top w} \Leftrightarrow \eta(x) = \frac{1}{1+e^{-x^\top w}}$$

Let's define the *sigmoid* function $\sigma : \mathbb{R} \to [0,1]$ as

$$\sigma(z) = \frac{1}{1+e^{-z}}.$$

which satisfies $\eta(x) = \sigma(x^\top w)$. Then the conditional distribution is given by

$$P(Y = y|X = x) = \sigma(x^\top w)^y \sigma(-x^\top w)^{1-y}.$$

Derivation of the maximum likelihood Given an i.i.d. training set, its empirical risk $\hat{\mathcal{R}}(w) = -\frac{1}{n}\log L(w)$ is

$$\begin{aligned}
\hat{\mathcal{R}}(w) &= -\frac{1}{n}\sum_{i=1}^n \Big\{y_i \log\sigma(x_i^\top w) - (1-y_i)\sigma(-x_i^\top w)\Big\} \\
&= \frac{1}{n}\sum_{i|y_i=1} \log(1+e^{-x_i^\top w}) + \frac{1}{n}\sum_{i|y_i=0} \log(1+e^{x_i^\top w}) \\
&= \frac{1}{n}\sum_{i=1}^n \log(1+e^{-\tilde{y}_i x_i^\top w})
\end{aligned}$$

where $\tilde{y}_i = 1$ if $y_i = 1$ and $\tilde{y}_i = -1$ if $y_i = 0$. It follows that maximum likelihood under this model is equivalent to empirical risk minimization for the *logistic loss* $\ell(\tilde{y}, z) = \log(1+e^{-\tilde{y}z})$ over the class of linear models.

Exercice: what is the maximum likelihood estimator for the model where we assume $P(X|Y = 1) \sim \mathcal{N}(\mu_1, I_d)$ and $P(X|Y = 0) \sim \mathcal{N}(\mu_0, I_d)$?

4 Convex surrogate losses for classification

We consider a classification task where $\mathcal{Y} = \{-1,1\}$. We depart from the "model based" approach of the previous section and only assume that the training data are i.i.d. samples from a random variable $(X, Y) \sim P$.

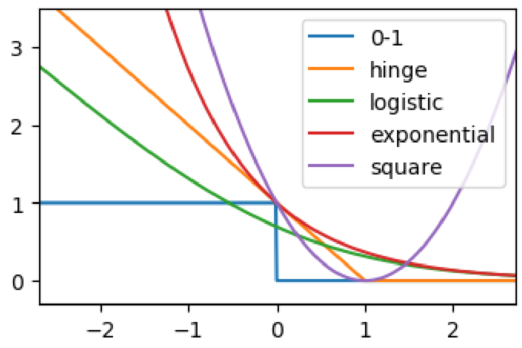

Figure 2: Classical losses used in classification tasks

- The relevant loss is the *0-1-loss* $\ell(y,z) = 1_{y \neq z}$ and, for a prediction function $f : \mathcal{X} \to \mathcal{Y}$ the risk $\mathcal{R}(f) = \mathbb{E}[\ell(Y, \text{sign}(f(X)))] = P(\text{sign}(f(X)) \neq Y)$ is the quantity that we truly want to minimize.
- However, the empirical risk $\hat{\mathcal{R}}(f)$ corresponds to minimizing the number of misclassifications, which is typically a difficult combinatorial optimization problem. We will see in the next lecture that we have efficient algorithms to minimize empirical risks if the loss is *convex* (i.e. always above its tangents).
- We thus consider *convex surrogate* losses. Denoting $\ell(y,z) = \phi(yz)$, commonly used examples include (see Figure 2 for illustrations):
  - the hinge loss $\phi_{hinge}(u) = (1-u)_+$ (the corresponding ℓ_2-regularized empirical risk minimization is called a "support vector machine"),
  - the logistic loss $\phi_{logistic}(u) = \log(1 + \exp(-u))$,
  - the square loss $\phi_{square}(u) = (1-u)^2$,
  - the exponential loss $\phi_{exp}(u) = \exp(-u)$.
- Let us denote by $\mathcal{R}_\phi$ and $\hat{\mathcal{R}}_\phi$ the associated (empirical or population) risk. The true risks with the 0-1-loss are denoted $\mathcal{R}$ and $\hat{\mathcal{R}}$. We minimize $\hat{\mathcal{R}}_\phi$ while the true goal is to minimize $\mathcal{R}$. What is the statistical cost of this computational convenience?

First, we can observe that $\ell(y, f(x)) \leq c\phi(yf(x))$ implies that $\mathcal{R}(f) \leq c\mathcal{R}_\phi(f)$. So a small $\mathcal{R}_\phi(f)$ gives a small $\mathcal{R}(f)$. But this is a rather weak assurance if, for example, $\inf_f \mathcal{R}_\phi(f) > 0$. For which choices of ϕ can we translate a control on $\mathcal{R}_\phi - \mathcal{R}_\phi^*$ into a control on $\mathcal{R} - \mathcal{R}^*$?

4.1 Classification calibration

For $x \in \mathcal{X}$, we define $\eta(x) = P(Y = 1|X = x)$. Then $\mathcal{R}_\phi(f) = \mathbb{E}[\phi(Yf(X))] = \mathbb{E}[\mathbb{E}[\phi(Yf(X))|X]]$, and

$$\begin{aligned}
\mathbb{E}\Big[\phi(Yf(X))|X = x\Big] &= P(Y = 1|X = x)\phi(f(x)) + P(Y = -1|X = x)\phi(-f(x)) \\
&= \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).
\end{aligned}$$

Define the conditional expectation and its optimal value:

$$C_\eta(\alpha) := \eta\phi(\alpha) + (1-\eta)\phi(-\alpha), \qquad\qquad H(\eta) := \inf_{\alpha\in\mathbb{R}} C_\eta(\alpha).$$

We have by then

$$\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi = \mathbb{E}\Big[C_{\eta(X)}(f(X))\Big] - \mathbb{E}\Big[H(\eta(X))\Big].$$

The prediction with minimal conditional risk is $f^*(x) = \operatorname{sign}(2\eta(x)-1)$. If the optimal conditional expectation $\mathbb{E}[\phi(Yf(X))|X=x]$ can be achieved with a value of α with the wrong sign, then minimizing $\mathcal{R}_\phi$ is not useful for classification. So define

$$H^-(\eta) := \inf\left\{C_\eta(\alpha)\ ;\ \alpha(2\eta-1)\le 0\right\}.$$

Definition 3 *We say that ϕ is* classification-calibrated *if, for all $\eta\neq 1/2$, $H^-(\eta) > H(\eta)$.*

In words, a loss is classification-calibrated iff its conditional expectation is minimized for a classifier of the same sign than the Bayes classifier.

Proposition 2 *For ϕ convex, ϕ is classification-calibrated iff ϕ is differentiable at 0 and $\phi'(0) < 0$. In this case, it holds $H^-(\eta) = \phi(0)$ for all $\eta \in [0,1]$ and $H(1/2) = \phi(0)$.*

Proof We only prove the "if" part (the "only if" is left as an exercise, or see [1]). For all $\eta\in[0,1]$, the function C_η is convex and satisfies $C'_\eta(0) = (2\eta-1)\phi'(0)$. Since a convex function lies above its tangents, if $2\eta-1>0$, this implies that C_η attains its minimum for $\alpha\in\,]0,+\infty]$ thus $H^-(\eta) > H(\eta)$. Similarly if $2\eta-1<0$, the minimizer is negative and $H^-(\eta) > H(\eta)$.

Also the fact that $C_\eta(\alpha) \geq C'_\eta(0)\alpha + C_\eta(0)$ implies that $H^-(\eta) = \phi(0)$. Finally, since by convexity $\phi(\alpha)/2 + \phi(-\alpha)/2 \geq \phi(0)$, it holds $H(1/2) = \phi(0)$ (taking $\alpha = 0$). ■

4.2 Bound on the true excess risk

Theorem 1 *Consider a nonnegative, convex and classification-calibrated ϕ. For any prediction function $f : \mathcal{X}\to\mathbb{R}$ and probability distribution P on $\mathcal{X}\times\{-1,1\}$, it holds*

$$\psi(\mathcal{R}(f) - \mathcal{R}^*) \leq \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi.$$

where $\psi(\theta) := H^-((1+\theta)/2) - H((1+\theta)/2) = \phi(0) - H((1+\theta)/2)$.

- Under those assumptions, we have $\lim_{\theta\to 0}\psi(\theta) = 0$ because H is continuous and $H(1/2) = \phi(0)$. This result implies that the true $0-1$ risk $\mathcal{R}(f)$ is controlled as soon as ϕ is convex, differentiable at 0 with $\phi'(0) < 0$. This is satisfied by all losses in Figure 2.
- We have the following values (for $\theta\geq 0$) for the losses defined above:
  - $\psi_{hinge}(\theta) = \theta$
  - $\psi_{logistic}(\theta) \geq \theta^2/2$

  - $\psi_{square}(\theta) = \theta^2$
  - $\psi_{exp}(\theta) = 1 - \sqrt{1-\theta^2} \sim_0 \theta^2/2$

- This gives some criterion to choose a loss for classification. In contrast to the maximum likelihood approach, it is not based on modeling but rather on performance guarantees. Note that the losses should satisfy additionnal properties to allow control of the estimation and optimization error.

Proof Using results from Lecture 1, we have, with $f^*(x) = \mathrm{sign}(\eta(x) - 1/2)$,

$$\mathcal{R}(f) - \mathcal{R}^* = \mathbb{E}\Big[1_{\{\mathrm{sign}(f(X)) \neq f^*(X)\}} |2\eta(X) - 1|\Big] = \mathbb{E}\Big[1_{\{f(X)(2\eta(X)-1) \leq 0\}} |2\eta(X) - 1|\Big].$$

Also, observe that H is concave (as an infimum of affine functions) and thus ψ is convex. Moreover, by the definition of H and H^-, $\psi(\theta) = \psi(-\theta)$. It follows that

$$\begin{aligned}
\psi\left(\mathcal{R}(f) - \mathcal{R}^*\right) &\leq \mathbb{E}\Big[\psi\Big(1_{\{\dots\}}|2\eta(X) - 1|\Big)\Big] \quad \text{(Jensen's inequality)} \\
&= \mathbb{E}\Big[1_{\{\dots\}}\psi(|2\eta(X) - 1|)\Big] \quad \text{(since } \psi(0) = 0\text{)} \\
&= \mathbb{E}\Big[1_{\{\dots\}}\psi(2\eta(X) - 1)\Big] \quad \text{(since } \psi(\theta) = \psi(-\theta)\text{)} \\
&= \mathbb{E}\Big[1_{\{\dots\}}(H^-(\eta(X)) - H(\eta(X)))\Big] \quad \text{(by the definition of } \psi\text{)} \\
&= \mathbb{E}\Big[1_{\{\dots\}}(\inf_{\alpha\,;\,\alpha(2\eta(X)-1)\leq 0} C_{\eta(X)}(\alpha) - H(\eta(X)))\Big] \quad \text{(by the definition of } H^-\text{)} \\
&\leq \mathbb{E}\Big[C_{\eta(X)}(f(X)) - H(\eta(X))\Big] \\
&= \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi.
\end{aligned}$$

■

References

[1] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.