Correct loss functions for generative models of supervised learning

Hông Vân Lê
Institute of Mathematics, Czech Academy of Sciences

Machine Learning, Optimization and Manifolds
Clermont-Ferrand, December 21, 2023

OUTLINE

1. Supervised learning and supervised operator.

2. Generative models of supervised learning and correct loss functions.

3. A variant of Vapnik-Stefanyuk's method of proving the learnability and its consequence.

4. Final remarks.

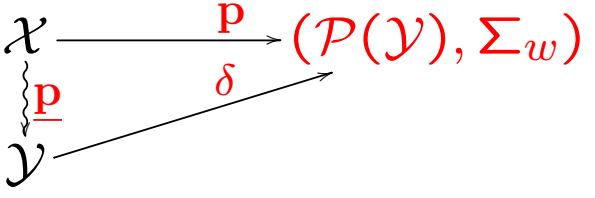
1. Supervised learning and supervised operator.

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces. In supervised learning, we are given a data set of labeled items:

$$S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in ((\mathcal{X} \times \mathcal{Y})^n, \mu^n),$$

where μ is an (unknown) probability measure such that $(x_n, y_n) \in (\mathcal{X} \times \mathcal{Y}, \mu)$. The goal of a learner in this scenario is to find the best approximation h_{S_n} of the conditional probability measure $[\mu_{\mathcal{Y}|\mathcal{X}}]$. We refer to $[\mu_{\mathcal{Y}|\mathcal{X}}]$ as the supervisor operator.

• $\mathbf{Probm}(\mathcal{X}, \mathcal{Y}) = \mathbf{Meas}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ - the set of all Markov kernels $\mathbf{p} : \mathcal{X} \times \Sigma_{\mathcal{Y}} \to [0, 1]$ s.t. (1) $\mathbf{p}(x, \cdot) \in \mathcal{P}(\mathcal{Y})$, (2) $\forall A \in \Sigma_{\mathcal{Y}}$, $\mathbf{p}(\cdot, A) : \mathcal{X} \to \mathbf{R}_{\geq 0}$ is measurable.



Σ_w is the smallest σ-algebra s.t. $\forall A \in \Sigma_{\mathcal{Y}}$ $I_A : \mathcal{P}(\mathcal{Y}) \to \mathbf{R}$, $\mu \mapsto \mu(A)$, is measurable.
$\delta(x) := \delta_x$.

• $\mathbf{p}$ is measurable and $\underline{\mathbf{p}}$ is a morphism. (Lawvere 1962).

- $\mathbf{Meas}(\mathcal{X},\mathcal{Y}) \subset \mathbf{Probm}(\mathcal{X},\mathcal{Y}) : f \mapsto \delta \circ f.$

- Regular conditional probability measures $\mu_{\mathcal{X}|\mathcal{Y}} \in \mathbf{Probm}(\mathcal{X},\mathcal{Y})$.

- For $T \in \mathbf{Probm}(\mathcal{X},\mathcal{Y})$ and $\mu \in \mathcal{M}(\mathcal{X})$ we let $T_*\mu \in \mathcal{M}(\mathcal{Y})$ s,.t. $\forall B \in \Sigma_{\mathcal{Y}}$

$$T_*\mu(A) = \int_{\mathcal{X}} \overline{T}(x)(A), d\mu(x).$$

- For $T \in \mathbf{Probm}(\mathcal{X},\mathcal{Y})$ the graph $\Gamma_T \in \mathbf{Probm}(\mathcal{X},\mathcal{X}\times\mathcal{Y})$ is generated by

$$\overline{\Gamma_T} : \mathcal{X} \to \mathcal{P}(\mathcal{X}\times\mathcal{Y}),\ x \mapsto \delta_x \cdot \overline{T}(x).$$

Theorem 1 (L. 2023)[Characterization of regular conditional probability measure]
A probabilistic morphism $T \in \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$ is a regular conditional probability measure of $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with respect to $\Pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y}$ if and only if

$$(\Gamma_T)_* \mu_{\mathcal{X}} = \mu$$

where $\mu_{\mathcal{X}} = (\Pi_{\mathcal{X}})_* \mu$.

2. Generative models of supervised learning and correct loss functions.

A generative model of supervised learning is a quintuple $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, R, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$, where $\mathcal{H} \subset \mathbf{Meas}(\mathcal{X}, (\mathcal{P}(\mathcal{Y}), \Sigma_w))$, $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \subset \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ - a set of all possible probability measures μ s.t. i.i.d. $(x, y) \in (\mathcal{X}\times\mathcal{Y}, \mu)$, and $R : \mathcal{H} \times \Big(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \cup \mathcal{P}_{emp}(\mathcal{X}\times\mathcal{Y})\Big) \to \mathbf{R} \cup \{+\infty\}$ is a risk/loss function, whose minimizer of $R_\mu := R(\mu, \cdot)$ on $\mathcal{H}$ are optimal predictors in the case μ is a distribution of observable data.

The aim of a learner is to find a "successful" algorithm

$$A : \cup_{i=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H},\ S_n \mapsto h_{S_n}.$$

• In order to find a successful algorithm, it is important to know that a loss function $R_\mu : \mathcal{H} \to \mathbf{R}$ measures the **deviation** of a predictor $h \in \mathcal{H}$ from the supervisor operator $\mu_{\mathcal{Y}|\mathcal{X}}$.

• A loss function $R : \mathcal{H} \times \Big(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \cup \mathcal{P}_{emp}(\mathcal{X} \times \mathcal{Y})\Big) \to \mathbf{R} \cup \{+\infty\}$ will be called $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$-correct, if $\exists\ \widetilde{\mathcal{H}} \subset \mathbf{Meas}(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ such that:

(1) $\mathcal{H} \subset \widetilde{\mathcal{H}}$.

(2) $\forall \mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}\ \exists\ h \in \widetilde{\mathcal{H}}$ s.t. $h = \mu_{\mathcal{Y}|\mathcal{X}}$.

(3) R is the restriction of a loss function $\widetilde{R} : \widetilde{\mathcal{H}} \times \Big(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \cup \mathcal{P}_{emp}(\mathcal{X} \times \mathcal{Y})\Big) \to \mathbf{R} \cup \{+\infty\}$ such that for any $\mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$

$$\arg\min_{h\in\widetilde{\mathcal{H}}} \widetilde{R}(h, \mu) = \{h \in \widetilde{\mathcal{H}} |\ [h]_{\mu_{\mathcal{X}}} = [\mu_{\mathcal{Y}|\mathcal{X}}]\}.$$

A loss function $R : \mathcal{H} \times \Big(\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \cup \mathcal{P}_{emp}(\mathcal{X} \times \mathcal{Y})\Big) \to \mathbf{R} \cup \{+\infty\}$ will be called correct, if R is the restriction of a $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$-correct loss function $\widetilde{R} : \mathcal{H} \times \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbf{R} \cup \{+\infty\}$.

• To define a correct loss function we need a functional characterization of $\mu_{\mathcal{Y}|\mathcal{X}}$, e.g. Theorem 1.

• Let $d : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \times \mathcal{P}(\mathcal{X}, \mathcal{Y}) \to \mathbf{R}_{\geq 0} \cup \{\infty\}$ be an arbitrary divergence. Then

$$R^d : \mathbf{Probm}(\mathcal{X}, \mathcal{Y}) \times \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbf{R}_{\geq 0},$$

$$R^d(h, \mu) = d((\Gamma_h)_* \mu_{\mathcal{X}}, \mu)$$

is a correct loss function by Theorem 1. For instance, Kullback-Leiber divergence, square loss function, and loss functions considered by Vapnik-Stefanuyk, which we shall consider next, are correct loss functions of this form.

3. A variant of Vapnik-Stefanyuk's method of proving the learnability and its consequence

• (VS) $\mathcal{X} \subset \mathbf{R}^n$, $\mathcal{Y} = \{w_1, \ldots, w_k\}$, $h \in \mathcal{H} \subset \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$. For $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $\mu_{\mathcal{X}} := (\Pi_{\mathcal{X}})_* \mu$, $i \in [1, k]$, for $v = (v_1, \ldots, v_n) \in \mathbf{R}^n$, let

$$F_\mu(v, \omega_i) := \mu([-\infty, v_1] \times \ldots \times [-\infty, v_n] \times \{w_i\}).$$

Letting $h(w_i|x) := h(x)(w_i)$, Theorem 1 implies

$$\int_{-\infty}^{v} h(\omega_i|x) d\mu_{\mathcal{X}}(x) = F_\mu(v, \omega_i) \; \forall i \in [1, k]$$

$$\overset{(3)}{\Longleftrightarrow} h = \mu_{\mathcal{Y}|\mathcal{X}}.$$

• Another example of Theorem 1: a function $p : I \to \mathbf{R}$ is a density function of a probability measure $\mu \in \mathcal{P}(I)$ if and only if

$$\int_{-\infty}^{x} p(t)dt = F_\mu(x) \, \forall x \in \mathbf{R}.$$

• Let us consider a general form of these equations for conditional probability measure $f \in \mathcal{H} \subset \mathbf{Probm}(\mathcal{X}, \mathcal{Y})$ for $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

$(*)$. $$A_\mu f = F_\mu$$

where F_μ depends on μ and needs not to be the cummulative distribution function.

• In many cases $f \in (E_1, d_1)$, $F_\mu \in (E_2, d_2)$ A_μ is a continuous operator and the equation (*) is ill-posed, i.e. the solution of (*) violates at least one of the 3 conditions of well-posedness in Hadamard sense: exists, unique, stable (the inverse operator A_μ^{-1} is continuous). Methods of solving ill-posed problems were proposed by Tikhonov (variational/regularization method, 1943,1962). Let us consider

$$(Af = F) \Longleftrightarrow f \in \arg\min_{f \in E_1} R^0(f)$$

We consider variational perturbed unstable equations

$$R_{\gamma(\varepsilon)}(f) = d^2_{E_2}(Af, F_\varepsilon) + \gamma(\varepsilon)W(f).$$

$$W(f) \geq 0 \,\&\, W^{-1}(c) \text{ compact } \forall c \in R^+.$$

Theorem (Tikhonov) If $\lim_{\varepsilon\to 0}\gamma(\varepsilon) = 0$ and $\lim_{\varepsilon\to 0}\frac{\varepsilon^2}{\gamma(\varepsilon)} = 0$ then

$$\lim_{\varepsilon\to 0} f_\varepsilon = f,$$

$$f_\varepsilon \in \arg\min_{f\in E_1} R_{\gamma(\varepsilon)}(f).$$

Stochastic ill-posed problem (Vapnik and Stefanuyk 1978-1998 for density, conditional density estimation and estimation of $\mu_{\mathcal{Y}|\mathcal{X}}$ if $\#(\mathcal{Y}) < \infty$).

F_μ and A_μ in Eq. $(A_\mu f = F_\mu)$ are unknown but $\forall\ \varepsilon > 0$

$$\lim_{n\to\infty} \mu^n\{S_n \in \mathcal{X}_n : d_E(F_{\mu_{S_n}}, F_\mu) > \varepsilon\} = 0.$$

We need to measure the closedness of operators $A_\mu, A_{\mu_{S_n}} : E_1 \to E_2$.

$$\|A_\mu - A_{\mu'}\| := \sup_{f\in E_1} \frac{\|A + \mu f - A_{\mu'} f\|_{E_2}}{W^{1/2}(f)}.$$

•Theorem 2.(L. 2023) (A variant of Vapnik-Stefanyuk's theorem). Let f_{S_l} be a γ_l^2-minimizer of $R^*_{\gamma_l}$ and f the solution of $Af = F$, where

$$R^*_{\gamma_l}(\widehat{f}, F_{S_l}, A_{S_l}) = d^2_E(A_{S_l}\widehat{f}, F_{S_l}) + \gamma_l W(\widehat{f}).$$

$\forall\ \varepsilon > 0$, $C_1, C_2 > 0$ $\exists\ \gamma_0 > 0$ s.t. $\forall\ \gamma_l \leq \gamma_0$:

$$\begin{gathered}(\mu_l)^*\{S_l \in \mathcal{X}_l :\ \rho_1(f_{S_l}, f) > \varepsilon\} \leq \\ (\mu_l)^*\{S_l \in \mathcal{X}_l :\ \rho_2(F_{S_l}, F) > C_1\sqrt{\gamma_l}\} \\ +(\mu_l)^*\{S_l \in \mathcal{X}_l :\ \|A_{S_l} - A\| > C_2\sqrt{\gamma_l}\}. \quad (1)\end{gathered}$$

• We need to use outer measure and convergence in outer probability, since it is complicated and some time impossible to choose a measurable learning algorithm which makes sense of all involved formulas with measures in theory of consistency of learning algorithms. Furthermore, some classical formulas also use the fact that uncountable of measurable subsets is measurable.

• To apply this theorem need to find W and estimate the second term in (2).

Examples of learnable overparameterized models of supervised learning (L.-2023).

• $\mathcal{X} = [0,1]^n \subset \mathbf{R}^n \times \{0\} \subset \mathbf{R}^{n+m}$, $\mathcal{Y} = [0,1]^m \subset \{0\} \times \mathbf{R}^m \subset \mathbf{R}^{n+m}$.

• Let $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ consist of probability measures $\mu_f := f\lambda_{n+m}$, where λ_{n+m} - the Lebesgue measure on $\mathbf{R}^{n+m}$, $f \in C^1(\mathcal{X}\times\mathcal{Y})$ and moreover $f(x,y) > 0$ for all $x, y \in \mathcal{X} \times \mathcal{Y}$.

$K : \mathbf{R}^{n+m}\times\mathbf{R}^{n+m} \to \mathbf{R}, (x,y) \mapsto \exp(-\|x-y\|_2^2)$.

• $\mathcal{H}_1 = C_{Lip}(\mathcal{X}, \mathcal{P}(\mathcal{Y})_{\tilde{K}})$ - the space of Lipschitz map to the metric space $\mathcal{P}(\mathcal{Y})_{\tilde{K}}$ whose metric $\tilde{K}$ is induced via the kernel mean embedding $\mathfrak{M}_K : \mathcal{P}(\mathcal{Y}) \to \mathcal{H}(K)$.

- $\mathfrak{M}_K(\mu) = \int_{\mathcal{Y}} K(y, \cdot) d\mu(y) \in \mathcal{H}(K)$.

- $\mathcal{H}_2$- a subspace in $C_{Lip}(\mathcal{X}, \mathcal{Y})$, which is embedded in $\mathcal{H}_1$ via the composition with the Dirac map $\delta : \mathcal{Y} \to \mathcal{P}(\mathcal{Y})$.

- $R_K : \mathcal{H} \times \mathcal{P}_{\mathcal{X} \times \mathcal{Y}}$, $(h, \mu) \mapsto \|(\Gamma_h)_* \mu_{\mathcal{X}} - \mu\|_{\tilde{K}}$.

• Gaussian kernel K has the following properties.

(i) C^1-differentiable.

(ii) the induced metric $\tilde{K}$ via $\mathfrak{M}_K$ defines the weak topology on $\mathcal{P}(\mathcal{Y})$ (Sriperumbudur 2016),

(iii) good rate of convergence in probability of $\mu_{S_n} \to \mu$ (LopezPas-Muandet-Schölkopf-Tolstikhin 2015).

• We can use and PDS kernel $K : \mathbf{R}^{n+m} \times \mathbf{R}^{n+m} \to \mathbf{R}$ with properties (i)- (iii) .

Theorem 3. (L. 2023) For $i = 1, 2$, the supervised learning model $(\mathcal{X}, \mathcal{Y}, \mathcal{H}_i, R_K, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ is learnable using the algorithm in Lê' variant of Vapnik-Stefanyuk's theorem.

• Vapnik applied their theorem to finite dimensional space of conditional densities on $\mathbf{R}$.

• $(\mathcal{X}, \mathcal{Y}, \mathcal{H}_i, \mathcal{P}_{\mathcal{X}\times\mathcal{Y}})$ is the first "overparameterized" model of supervised learning which is learnable.

5. Discussion of results

• Our generative model of supervised learning (including characterizations of regular conditional probability measures) encompasses main problems of statistical inferences and offers unifying and broader perspectives.

• Almost all classical algorithms for density estimations can be obtained by using standard regularization method of solving stochastic problems $Af = F$ where A is fixed and F is 'stochastically'' approximated.

• We have several different characterizations of regular conditional probability measures and using some of them we can extend Cucker-Smale theorem for (1) regression problem in Euclidean spaces to regression problem in Hilbert spaces, and (2) to a problem of estimating conditional probability where input and label spaces are sitting in Euclidean spaces.

• Let us consider Bayesian statistical model $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ where $\mathbf{p} \in \mathbf{Probm}(\Theta, \mathcal{X})$. Then the marginal (predictive) probability $\mu_\mathcal{X} = \mathbf{p})_* \mu_\Theta$.

Theorem 4.(Jost-L.-Tran, 2021) $\mathbf{q} : \mathcal{X} \to \mathcal{P}(\Theta)$ is a Bayesian inversion of $\mathbf{p}$ relative to μ_Θ i.e. a regular conditional probability measure of the joint distribution $(\Gamma_\mathbf{p})_* \mu_\Theta$ w.r.t. $\Pi_\mathcal{X} : \Theta \times \mathcal{X} \to \mathcal{X}$ if and only if

$$\mathbf{q}_* \mathbf{p}_* \mu_\Theta = \mu_\Theta.$$

• Using Theorem 4, and motivated by the concept of a correct loss function, we can develop general Bayesian decision models encompassing concepts of Bayesian inference, classical Bayesian decision theory, which serve as models for Bayesian methods for density estimations and regression problems.

References.

[1] V. Vapnik, Statistical Learning Theory. John Willey & Sons, 1998.

[2] V. Vapnik, The Nature of Statistical Learning Theory. Springer, 2nd Edition, 2000.

[3] V. Vapnik and A. Stefanyuk, Nonparametric methods for estimating probability densities, Automation and Remote Control, 8 (1978), 38-52.

[4] V. Vapnik and R. Izmailov, Rethinking statistical learning theory: learning using statistical invariants, Machine Learning (2019) 108:381-423.

[5] F. Cucker and S. Smale, On mathematical foundations of learning, Bulletin of AMS, 39 (2002), 1-49.

[6] J. Jost, H. V. Lê, and T. D. Tran, Probabilistic morphisms and Bayesian nonparametrics, Eur. Phys. J. Plus 136, 441 (2021), arXiv:1905.11448.

[7] H. V. Lê, Supervised learning with probabilistic morphisms and kernel mean embeddings, arXiv:2305.06348.

[8] H. V. Lê, H. Q. Minh, F. Protin, W. Tuschmann, Mathematical Foundations of Machine Learning, in preparation, (Springer Nature, 2025?).

[9] B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G.R.G. Lanckriet, Hilbert space embeddings and metrics on probability measures, J. Math. Learn. Res. 11, 1517-1561, (2010).

[10] B. Sriperumbudur, On the optimal estimation of probability measures in weak and strong topologies, Bernoulli, 22(3):1839-1893, 08, 2016.

THANK YOU FOR YOUR ATTENTION!