Machine Learning & Deep Learning

Tutorial

MSCV/ESIREM

Antoine Lavault
antoine.lavault@u-bourgogne.fr

Linear Classifiers and friends

Any question or exercise marked with a "*" is typically more technical or goes further into developing the tools and notions seen during class.

Problem 1

Basic concepts.

1. Given a vector space X, a norm on X is a real-valued function $p : X \to \mathbb{R}$ with the following properties, where $|s|$ denotes the usual absolute value of a scalar s:

   - Subadditivity/Triangle inequality: $p(x + y) \leq p(x) + p(y), \forall x, y \in X$
   - Absolute homogeneity: $p(sx) = |s|p(x), \forall x \in X$ and all scalars s.
   - Positive definiteness: $\forall x \in X p(x) = 0 \Rightarrow x = 0$.

   Show that the L_1, L_2, and infinity norms are indeed norms. For the sake of simplicity, we will assume X to be a finite-dimensional space.

2. Why is a linear classifier generally too weak for most classification tasks?

3. Is the logistic regression a linear classifier?

1. K-Means will find k-clusters, a local minimum of the sum of the squared distance between objects in the dataset and the centroid of the cluster they belong to.

2. Real-world data exhibits complex and non-linear relationships between features and class labels. Linear classifiers cannot capture these intricate patterns, which makes them underperform on tasks where the decision boundary is not a hyperplane. Note that it is possible to "linearize" the features to some extent, but this is outperformed by non-linear methods such as kernel-based SVMs or neural networks. Note that the curse of dimensionality is not a problem, as things are generally "more separable" in higher dimensional spaces, which is the foundation for kernel-based SVMs.

3. Yes, it is. The logistic regression uses a linear combination of the input features, then calculates the weighted sum of the input features and applies a logistic (sigmoid) function to this linear combination. The result is a probability score that indicates the likelihood of belonging to a particular class. The decision boundary is also a hyperplane.

Problem 2

Gradient Descent Mechanics. Gradient descent is the primary algorithm to search optimal parameters for our ML and DL models. Typically, we want to solve optimization problems stated as

$$\min_{\theta \in \Theta} L(f_\theta, \mathcal{D}),$$

where $\mathcal{L}$ are differentiable functions. In this example, we look at a simple supervised learning problem where given a dataset $\mathcal{D} = \{(x_i, y_i)\}_N$, we want to find the optimal parameters θ that minimize some loss function. We will consider different models for learning the mapping from input to output and examine the behavior of gradient descent for each model.

1. The simplest parametric model entails learning a single-parameter constant function. We wish to find

$$\hat{\theta}_{const} = \min_{\theta \in \mathbb{R}} \mathcal{L}(f_\theta, \mathcal{D}) = \min_{\theta \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta)^2$$

   (a) What is the gradient of L w.r.t. θ? (w.r.t. means "with respect to").

   (b) What is the optimal value of θ?

   (c) Write the gradient descent update rule.

   (d) Stochastic Gradient Descent (SGD) is an alternative optimization algorithm where instead of using all N samples, we use a single sample per optimization step to update the model. What is the gradient update in that case? Assuming we sample uniformly, what is the contribution of each data point to the full gradient update (do the sum of the updates)?

   Note: this 1-sample-only rule only simplifies the calculations. In general, the SGD is used on batches of n samples.

2. Instead of constant functions, we now consider a single-parameter linear model $\hat{y}(x_i) = \theta x_i$, where we search for θ such that:

$$\hat{\theta} = \min_{\theta \in \mathbb{R}} \mathcal{L}(f_\theta, \mathcal{D}) = \min_{\theta \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta x_i)^2$$

   (a) What is the gradient of L w.r.t. θ?

   (b) What is the optimal value of θ?

   (c) Write the gradient descent update rule.

   (d) Do all points get the same "weight" in the update? Why or why not?

1. First, $\partial L/\partial\theta = 2/N \sum_{i=1}^{N} (\theta - y_i)$. The optimality condition can be expressed as $\partial L/\partial\theta = 0$, which gives us $\theta^* = 1/N \sum_{i=1}^{N} y_i$.

   The gradient update rule can be expressed as $\theta_{t+1} \leftarrow \theta_t + \alpha \partial L/\partial\theta$.

   With SGD, it does not change anything for the whole gradient (do the sum over the y_i to check). The update is $\theta_{t+1} \leftarrow \theta_t + 2\alpha/N(\theta - y_i)$

2. First, $\partial L/\partial\theta = 2/N \sum_{i=1}^{N}(\theta x_i - y_i)x_i$. The optimality condition can be expressed as $\partial L/\partial\theta = 0$, which gives us $\theta^* = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$.

   The contribution of each data point to the full gradient depends on the magnitude of the data point x_i, i.e., data points closer to the origin contribute less to the full gradient.

Problem 3

Why choosing a learning rate is a pain in the GPU. To understand the role of the learning rate, it is useful to understand it in the context of the simplest possible problem first. Suppose we want to solve the $\sigma w = y$ scalar equation where $\sigma > 0$. We proceed with an initial condition $w_0 = 0$ by using gradient descent to minimize a squared loss error.

1. Write the loss function and its derivative with respect to w.

   $L(w) = (y - \sigma w)^2$ and $-2\sigma(y - \sigma w)$

2. Write the gradient descent update with a learning rate of η for this optimization problem. Present the result under the form $f(\eta, \sigma)w_t + g(\sigma, \eta, y)$.

   $w_{t+1} = w_t + 2\eta\sigma(y - \sigma w_t) = (1 - 2\eta\sigma^2)w_t + 2\eta\sigma y$

3. Show that $s_t = w_t - y/\sigma$ is a geometric progression. Deduce an expression for w_t. For what learning rate values $\eta > 0$ is the recurrence stable?

   First, $s_{t+1} = (1 - 2\eta\sigma^2)s_t$, which means $w_t = (1 - 2\eta\sigma^2)^t(w_0 - y/\sigma) + y\sigma$.

   Finally, the recurrence is stable if $|1 - 2\eta\sigma^2| < 1$, i.e., $\eta < 1/\sigma^2$.

4. The previous question gives us an upper bound for the learning rate η that depends on σ beyond which we cannot safely go. If η is below that upper bound, how fast does w_t converge to its final solution $w^* = y/\sigma$, i.e., if we wanted to get within a factor $(1 - \varepsilon)$ of w^*, how many iterations t would we need?

   This problem can be expressed as finding a lower bound for T such that $|w_T - w^*| < \varepsilon|w^*|$.

   We can first remark that $|w_T - w^*| = |s_T|$ which in turns gives us $|1 - 2\eta\sigma^2|^T < \varepsilon$.

   Finally, we have $T > \frac{\log \varepsilon}{\log|1-2\eta\sigma^2|}$

Problem 4

Information Theory and Classification (*) This exercise shows how it is possible to interpret the logistic regression as something other than likelihood.

Some Results on Shannon Entropy. Shannon entropy is a mathematical function developed by Claude Shannon. It measures the amount of information a given source contains or delivers. This source can take different forms, such as a text written in a specific language, an electrical signal, or even a computer file (a collection of bytes). Shannon entropy provides an intuitive measure of the uncertainty or unpredictability associated with a source of information. It quantifies the complexity and information richness of a data set. The higher the entropy, the more unpredictable and novel the information source.

For a source, which is a discrete random variable X comprising n symbols $(x_1, \cdots, x_n)$, each symbol x_i having a probability P_i of appearing, the entropy H of the source X is defined as :

$$H_b(X) = -\mathbb{E}[\log_b P(X)] = -\sum_{i=1}^{n} P_i \log_b P_i. \tag{1}$$

for a logarithm in base $b > 1$. In the following, the logarithm will be in base e (Napierian logarithm, corresponding to *nats*) and will be denoted $\log$.

1. Propose bounds of P_i and a condition on their sum.

   $0 \leq P_i \leq 1$ et $\sum P_i = 1$ car les P_i sont liés à des probabilités (discrètes).

2. Let L be the Lagrangian of the Shannon entropy-constrained maximization problem:

$$L(P_1, \cdots, P_n, \lambda) = \sum_{i=1}^{n} P_i \log_b P_i - \lambda(\sum P_i - 1) \tag{2}$$

   The following questions describe the process of calculating the Lagrangian to obtain the entropy-maximizing distribution of X.

   (a) Calculate the partial derivatives of L with respect to P_i. Equating the partial derivative to 0, derive an expression for P_i as a function of λ.

   $dL/dP_i = -\log P_i - 1 - \lambda = 0, \forall i$

   D'où : $P_i = \exp(-1 - \lambda), \forall i$

   Les P_i sont donc tous les mêmes.

   (b) Calculate the partial derivative of L with respect to λ and set the result to 0.

   On obtient : $\sum P_i - 1 = 0$. On retrouve la condition exprimée plus haut.

   (c) Using the results of the previous questions and remembering that solving the Lagrangian is like solving a constrained optimization problem, what is the distribution $\hat{X}$ that maximizes entropy?

   On sait que tous les P_i sont les mêmes. Donc: $1 = n \exp(-1 - \lambda)$.

   Finalement, $\lambda = \log(n - 1)$ et $P_i = 1/n$.

   Les P_i suivent donc une loi uniforme et maximise l'information.

   (d) Interpret the result.

   This result is not shocking: all the symbols in the alphabet have an equal chance of coming out. No symbol is less important than any other. However, such a code will be difficult to compress.

Entropy and Logistic Regression. Remember that in the case of logistic regression, the explained variable Y is a binary variable, which can represent a qualitative property. It can only take the values 0 or 1. The explanatory variables $X_1, \cdots, X_p$ are real, and are grouped as $X = (X_1, \cdots, X_p)$.

It is assumed that $Y \sim \text{Bernoulli}(p(X\beta))$ with β the parameters of the regression with $p(z) = \frac{1}{1+e^{-z}}$.

1. Show that $\log \frac{p(z)}{1-p(z)} = z$.

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

   D'où le résultat.

The output of the logistic regression model can be interpreted as a probability that the input belongs to one class or as a probability that it belongs to the other class in a binary classification problem. We denote this probability as follows:

$$P(Y = 1|z) = p(z)$$

1. What is the probability $P(Y = 0|z)$ for an observation z as a function of p?

   On remarquera que Y suit une loi de Bernouilli, donc : $P(Y = 0|z) = 1 - P(Y = 1|z) = 1 - p(z)$

2. The likelihood for an observation (x, y) is given by the probability $P(y|x, \beta)$. What are the values taken by $P(y|x, \beta)$ depending on the value of y as a function of p?

   Maximum likelihood for a parametric family θ is the estimator used in logistic regression and is generally given by :

   $$\underset{\theta}{\text{argmax}}\ \mathcal{L}_n(\theta) = \mathcal{L}_n(\theta, \mathbf{y}) = f_n(\mathbf{y}, \theta)\,, \tag{3}$$

   with

   $$f_n(\mathbf{y}, \theta) = \prod_{k=1}^{n} f_k^{\text{observation}}(y_k, \theta)\,.$$

   (a) Please define the likelihood for all obervations X.
   $\prod_{k=1}^{n} f_k^{\text{observation}}(y_k, \theta) = \prod_{k=1}^{n} P(Y = y_k|x_k, \beta)$

   (b) The optimization of a product is generally difficult. Propose a transformation to facilitate the maximum likelihood optimization. Justify.

   Le logarithme est une bonne idée pour transformer le produit en somme, ce qui rendra le processus d'optimisation plus aisé, en particulier en évitant la règle de différenciation d'un produit. On notera que le logarithme ne change pas le signe et ne pose pas de problème de définition.

3. Show that the result obtained in the previous question can be rewritten as :

   $$\mathrm{H}(p, q) = -\sum_{x} p(x)\, \log q(x). \tag{4}$$

   with p and q to define.

   On remarquera que $P(Y = y|x, \beta) = p(z)^y (1-p(z))^{1-y}$ et que $\log P = y \log(p) + (1-y) \log(1-p)$

   En choisissant $p = Y$ et $q = p(\cdot)$

4. Noting the similarity with equation 1, what interpretation could be given to the quantity $H(\cdot,\cdot)$? Infer the link between this quantity and logistic regression in terms of information.

   Note: $H(\cdot,\cdot)$ is called cross-entropy.

   $H(\cdot,\cdot)$ corresponds to the information that is not redundant between the two distributions.

   Logistic regression, therefore, conciliates two probability distributions by minimizing the information needed to differentiate the induced distributions.

Problem 5

Entropy, Cross-Entropy, Kullback-Leibler (KL)-divergence. Entropy is a fundamental concept in information theory and statistics, representing the measure of uncertainty or disorder in a system. It quantifies the unpredictability of outcomes in a given probability distribution. Higher entropy signifies greater randomness and less predictability, while lower entropy indicates more order and predictability. It is noted H and is given by:

$$H(Y) = E_Y[-\log p(Y=k)] = -\sum_k [p(Y=k)\log p(Y=k)]$$

On the other hand, cross-entropy is a concept closely related to entropy, often used in machine learning and statistics. It measures the dissimilarity between two probability distributions, typically predicted and true data distributions. Cross-entropy is a loss function in various machine learning tasks, such as classification, to guide model training by penalizing predictions that diverge from the true distribution. The cross-entropy is given by:

$$\mathrm{H}(p,q) = -\sum_x [p(x)\ \log q(x)].$$

Kullback-Leibler (KL) divergence is a mathematical measure of the difference between two probability distributions. Specifically, it quantifies how one distribution diverges from another. KL divergence is asymmetric and can be considered a way to measure the inefficiency of using one distribution to approximate another:

$$D_{KL}(p \parallel q) = \sum_x p(x) log[p(x)/q(x)]$$

Finally, a similar measure to the KL-Divergence is the Jensen-Shannon divergence, which is given by:

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2}D_{KL}(P \parallel M) + \frac{1}{2}D_{KL}(Q \parallel M),$$

where $M = \frac{1}{2}(P+Q)$ is a mixture distribution of P and Q.

1. Let's define two probability distributions given by $p(x)$, equals to 1 or -1 with probability 0.5, and $q(x)$ equals 1 with probability 0.1 and -1 with probability 0.9.

   Show the KL-divergence is not symmetric. What about the Jensen-Shannon divergence?

   $D(p \parallel q) = 0.5\log(0.5/0.1) + 0.5\log(0.5/0.9) \neq 0.1\log(0.1/0.5) + 0.9\log(0.9/0.5)$.

   The JS divergence is symmetric.

2. Show that the Kullback-Leibler divergence is non-negative and that it is equal to 0 when $p = q$. Is the Jensen-Shannon divergence non-negative as well?

   First, let's remark that $\log a \leq a - 1, \forall a > 0$. We then have:

$$-D_{KL}(p \parallel q) = \sum_x p(x) log[q(x)/p(x)] \leq \sum_x p(x)(q(x)/p(x) - 1)] = \sum q - \sum p = 0$$

   If $p = q$, the equality is trivial. For the other way around, note that $\log a < a - 1, \forall a > 0, a \neq 1$. This means the inequality is strict whenever $p/q \neq 1$, i.e., $p \neq q$. The equality only happens when and only when $p = q$.

3. Show that $D_{KL}(p \parallel q) = H(p, q) - H(p)$. Deduce another expression of the JS divergence, function of p, q, and M, the mixture distribution. Easy. Just write the things down. And that way, $JSD = H(M) - \frac{1}{2}(H(p) + H(q))$

4. Show that $E(\log \frac{2}{1+e^x}) \leq \log \frac{2}{1+e^{E(x)}}$.

   Hint: $\varphi(E[X]) \leq E[\varphi(X)]$. *when* φ *is convex. And* $f(x) = \log(1 + e^x)$ *is convex.*

   From the hint, $\log(1 + e^{E[X]}) \leq E[\log(1 + e^x)]$.. Which can then be inverted, hence the result.

5. Show that $D_{KL}(p, q) \leq \sum p^2/q - 1$

   $D_{KL}(p, q) = \sum_i p_i \log(p_i/q_i) \leq \sum_i p_i(p_i/q_i - 1) \sum_i p^2/q - \underbrace{\sum_i p_i}_{=1}$

6. Show the Jensen-Shannon divergence is bounded by 1.

   Put the results above together to bound $D_{KL}(p, (p + q)/2)$.

   We then have that $D_{KL}(p, (p + q)/2) \leq \sum p \frac{2}{\frac{p+q}{p}} - 1 = \sum p \frac{2}{1+q/p} - 1$

   We can then use question 4 while remarking that $\sum_i p \frac{2}{1+q/p} = E_p(\log \frac{2}{1+e^{\log q/p}})$.

   Finally, we have the upper bound $\frac{2}{1+e^{-D(p,q)}} - 1 \leq 1$ since the Kullback-Leibler divergence is non-negative.