Some machine learning algorithms

Maxime Sangnier

September 30, 2019

List of Algorithms

Chapter 1

Supervised learning

Algorithm 1 Adaboost.

Input: $T \in \mathbb{N}$ (number of iterations), $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ (training sample).
  for $i = 1$ **to** n **do**
    $D_1(i) \leftarrow \frac{1}{n}$
  end for
  $f_0 = 0$ *(null function)*
  for $t = 1$ **to** T **do**
    $g_t \leftarrow$ base $\{\pm 1\}$-classifier from $\mathcal{C}$ with small error $\epsilon_t = \sum_{i=1}^n D_t(i) \mathbf{1}_{Y_i \neq g_t(X_i)}$
    $w_t \leftarrow \arg\min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \exp\left(-Y_i(f_{t-1}(X_i) + w g_t(X_i))\right) = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ *(ERM)*
    $Z_t \leftarrow \sum_{i=1}^n D_t(i) \exp\left(-w_t Y_i g_t(X_i)\right) = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ *(normalization)*
    for $i = 1$ **to** n **do**
      $D_{t+1}(i) \leftarrow D_t(i) \exp\left(-w_t Y_i g_t(X_i)\right) / Z_t$
    end for
    $f_t = \sum_{j=1}^t w_j g_j$
  end for
Output: $g_n^T = \operatorname{sign}(f_T)$.

Algorithm 2 Gradient boosting.

Input: $T \in \mathbb{N}$ (number of iterations), $\nu \in (0, 1]$ (shrinkage coefficient), $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ (training sample).
  $f_0 \in \arg\min_{\gamma \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n L(Y_i, \gamma)$ *(constant function)*
  for $t = 1$ **to** T **do**
    for $i = 1$ **to** n **do**
      $r_{i,t} \leftarrow -\ell_i'(f_{t-1}(X_i))$ *(pseudo-residuals)*
    end for
    $g_t \leftarrow$ base regressor from $\mathcal{R}$ for the training set $\{(X_i, r_{i,t})\}_{1 \leq i \leq n}$
    $w_t \leftarrow \arg\min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n L(Y_i, f_{t-1}(X_i) + w g_t(X_i))$ *(line search)*
    $f_t = f_{t-1} + \nu w_t g_t$
  end for
Output: f_T.

Algorithm 3 Sequential minimal optimization.

Input: $C > 0$ (tradeoff parameter), $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (kernel function), $\{(X_i, Y_i)\}_{1 \le i \le n}$ (training sample).

$Q \leftarrow (k(X_i, X_j) Y_i Y_j)_{1 \le i,j \le n}$ *(labeled kernel matrix)*

while not converged **do**

find α_i for which Karush-Kuhn-Tucker (KKT) conditions are violated

pick $\alpha_j \neq \alpha_i$ at random

solve Problem **??** with respect to (α_i, α_j) with all other variables fixed

end while

Output: $(\alpha_1, \ldots, \alpha_n)$.

Chapter 2

Clustering

Algorithm 4 Sampling of a mixture model.

Input: $(P_{\theta_1}, \dots, P_{\theta_m})$ (mixture components) and $(\pi_1, \dots, \pi_m)$ (probability vector).
$z \leftarrow$ sample from $\mathcal{M}(1, \pi_1, \dots, \pi_m)$ *(multinomial variable)*
$y \leftarrow \sum_{j=1}^{m} j \mathbf{1}_{z_j=1}$ *(cluster label)*
$x \leftarrow$ sample from P_{θ_y}.
Output: x.

Algorithm 5 EM algorithm.

Input: $T \in \mathbb{N}$ (number of iterations), X (observed sample).
$\hat{\theta}_0 \leftarrow$ random initialization
for $t = 0$ **to** $T - 1$ **do**
set $Z^{(t)}|X \sim Q_{\hat{\theta}_t, X}$
E step: compute $F(\theta|\hat{\theta}_t) = \mathbb{E}\left[\log\left(g_\theta(X, Z^{(t)})\right)|X\right]$
M step: set $\hat{\theta}_{t+1} \in \arg\max_{\theta \in \Theta} F(\theta|\hat{\theta}_t)$
end for
Output: $\hat{\theta}_T$.

Algorithm 6 EM algorithm (maximization-maximization).

Input: $T \in \mathbb{N}$ (number of iterations), X (observed sample).
$\hat{\theta}_0 \leftarrow$ random initialization
for $t = 0$ **to** $T - 1$ **do**
E step: set $\hat{\theta}'_t \in \arg\max_{\theta \in \Theta} F(\hat{\theta}_t|\theta) + H(\theta)$, that is $\hat{\theta}'_t = \hat{\theta}_t$ *(best lower bound of* $\log(m_\theta(X))$ *knowing* $\hat{\theta}_t$*)*
M step: set $\hat{\theta}_{t+1} \in \arg\max_{\theta \in \Theta} F(\theta|\hat{\theta}_t)$ *(maximize the lower bound given* $\hat{\theta}_t$*)*
end for
Output: $\hat{\theta}_T$.

Algorithm 7 EM for Gaussian mixtures (soft k-means).

Input: $\{X_i\}_{1\le i\le n}$ (training sample).
  $\pi_j \leftarrow \frac{1}{k}$, for all $j \in [k]$ *(initialization)*
  $\mu_j \leftarrow$ random point, for all $j \in [k]$
  $\Sigma_j \leftarrow$ overall sample covariance, for all $j \in [k]$
  while not converged **do**
    $p_{ij} \leftarrow \frac{\pi_j \phi_{(\mu_j,\Sigma_j)}(X_i)}{\sum_{\ell=1}^k \pi_\ell \phi_{(\mu_\ell,\Sigma_\ell)}(X_i)} \approx \mathbb{P}(Y_i = j|X_i)$ *(expectation)*
    $\pi_j \leftarrow \frac{1}{n}\sum_{i=1}^n p_{ij}$ *(maximization)*
    $\mu_j \leftarrow \frac{\sum_{i=1}^n p_{ij} X_i}{\sum_{i=1}^n p_{ij}}$
    $\Sigma_j \leftarrow \frac{\sum_{i=1}^n p_{ij}\left[(X_i-\mu_j)(X_i-\mu_j)^\top\right]}{\sum_{i=1}^n p_{ij}}$
  end while

Algorithm 8 k-means.

Input: $T \in \mathbb{N}$ (number of iterations), $\{X_i\}_{1\le i\le n}$ (training sample).
  $\hat{\mu}_i \leftarrow$ random point from $\mathcal{X}$ for all $i \in [n]$ *(initialization)*
  for $t = 1$ **to** T **do**
    compute a Voronoi partitioning $(C_1, \dots, C_k)$ corresponding to cluster centers $(\hat{\mu}_1, \dots, \hat{\mu}_k)$
    $\hat{C}_j \leftarrow \{X_1, \dots, X_n\} \cap C_j$ for all $j \in [k]$
    $\hat{\mu}_j \leftarrow \frac{1}{|\hat{C}_j|}\sum_{X\in\hat{C}_j} X$
  end for
Output: $(C_1, \dots, C_k)$.

Algorithm 9 k-means++.

Input: $T \in \mathbb{N}$ (number of iterations), $\{X_i\}_{1\le i\le n}$ (training sample).
  $\hat{\mu}_1 \leftarrow$ random point from $\{X_i\}_{1\le i\le n}$ *(initialization)*
  for $j = 2$ **to** k **do**
    $\hat{\mu}_j \leftarrow$ random point from $\{X_i\}_{1\le i\le n}$ with density $\sum_{i=1}^n \frac{\Delta_j(\cdot)^2}{\sum_{\ell=1}^n \Delta_j(X_\ell)^2}\delta_{X_i}(\cdot)$
  end for
  $(C_1, \dots, C_k) \leftarrow$ output of k-means algorithm based on $(\hat{\mu}_1, \dots, \hat{\mu}_k)$
Output: $(C_1, \dots, C_k)$.

Algorithm 10 Unnormalized spectral clustering.

Input: $W \in \mathbb{R}^{n\times n}$ (adjacency matrix).
  $L \leftarrow$ Laplacian of W
  $H \leftarrow k$ minor eigenvectors of L as columns
  $Y_i \leftarrow i^{th}$ row of H (for all $i \in [n]$) *($Y_i \in \mathbb{R}^k$)*
  $(\hat{C}_1, \dots, \hat{C}_k) \leftarrow$ output of k-means algorithm based on $(Y_1, \dots, Y_n)$
Output: $(\hat{C}_1, \dots, \hat{C}_k)$.

Algorithm 11 Normalized spectral clustering (with L_w).

Input: $W \in \mathbb{R}^{n\times n}$ (adjacency matrix).
  $L_w \leftarrow$ Laplacian of W
  $H \leftarrow k$ minor eigenvectors of L_w as columns *(similar to the generalized eigenproblem $Lu = \lambda Du$)*
  $Y_i \leftarrow i^{th}$ row of H (for all $i \in [n]$) *($Y_i \in \mathbb{R}^k$)*
  $(\hat{C}_1, \dots, \hat{C}_k) \leftarrow$ output of k-means algorithm based on $(Y_1, \dots, Y_n)$
Output: $(\hat{C}_1, \dots, \hat{C}_k)$.

Algorithm 12 Normalized spectral clustering (with L_s).

Input: $W \in \mathbb{R}^{n \times n}$ (adjacency matrix).
$L_s \leftarrow$ Laplacian of W
$H \leftarrow k$ minor eigenvectors of L_s as columns
$Y_i \leftarrow i^{th}$ row of H normalized to 1 (for all $i \in [n]$) *($Y_i \in \mathbb{R}^k$, $\sum_{j=1}^k (Y_i)_j^2 = 1$)*
$(\hat{C}_1, \ldots, \hat{C}_k) \leftarrow$ output of k-means algorithm based on $(Y_1, \ldots, Y_n)$
Output: $(\hat{C}_1, \ldots, \hat{C}_k)$.

Algorithm 13 DBSCAN.

Input: $\epsilon > 0$ (neighborhood radius), $m \in \mathcal{N}$ (minimal number of neighbors), $\{X_i\}_{1 \leq i \leq n}$ (training sample).
$T \leftarrow \{X_i\}_{1 \leq i \leq n}$ *(unlabeled points)*
$k \leftarrow 0$ *(current number of clusters)*
while $T \neq \emptyset$ **do**
  pick X in T
  $N \leftarrow \epsilon$-neighborhood of X
  if $|N| \geq m$ **then**
    $k \leftarrow k + 1$
    initialize a new cluster $\hat{C}_k = \emptyset$
    move X from T to $\hat{C}_k$
    $S \leftarrow (N \backslash \{X\}) \cap T$ *(unlabeled neighbors)*
    while $S \neq \emptyset$ **do**
      pick Y in S
      move Y from S to $\hat{C}_k$ (and remove Y from T)
      $N' \leftarrow \epsilon$-neighborhood of Y
      if $|N'| \geq m$ **then**
        $S \leftarrow S \cup (N' \cap T)$ *(unlabeled neighbors)*
      end if
    end while
  end if
end while
Output: $(\hat{C}_1, \ldots, \hat{C}_k, T)$ *(k clusters and a set of outliers)*

Chapter 3

Dimensionality reduction

Algorithm 14 Reduced representation by PCA.

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$ (data matrix), p (reduced dimension).
  Second order matrix
  $C \leftarrow \mathbf{X}^\top \mathbf{X}$
  $V \leftarrow p$ major eigenvectors of C
  $U \leftarrow \mathbf{X}V$
  Gram matrix
  $K \leftarrow \mathbf{X}\mathbf{X}^\top$
  $\lambda_1, \ldots, \lambda_p \leftarrow p$ major eigenvalues
  $V \leftarrow p$ major eigenvectors of K
  $U \leftarrow \left[\sqrt{\lambda_1} v_1 | \ldots | \sqrt{\lambda_p} v_p\right]$
  singular value decomposition (SVD)
  $\lambda_1, \ldots, \lambda_p \leftarrow p$ major singular values of $\mathbf{X}$
  $V \leftarrow p$ major left singular vectors of $\mathbf{X}$
  $U \leftarrow [\lambda_1 v_1 | \ldots | \lambda_p v_p]$
Output: $U \in \mathbb{R}^{n \times p}$.

Algorithm 15 Classical multidimensional scaling.

Input: $D \in \mathbb{R}^{n \times n}$ (matrix of squared pairwise distances), $p \in [n]$ (reduced dimension).
  $K_{\mathbf{X}'} \leftarrow -\frac{1}{2} H D_{\mathbf{X}} H$
  Compute the eigendecomposition $\sum_{i=1}^{n} \lambda_i v_i v_i^\top$ of $K_{\mathbf{X}'}$, with $\lambda_1 \geq \cdots \geq \lambda_n$
  $\mathbf{Z} \leftarrow \left[\sqrt{\lambda_1} v_1 | \ldots | \sqrt{\lambda_p} v_p\right] \in \mathbb{R}^{n \times p}$
  $\{z_i\}_{1 \leq i \leq n} \leftarrow$ rows of $\mathbf{Z}$
Output: $\{z_i\}_{1 \leq i \leq n}$.

Algorithm 16 SMACOF.

Input: $d \in \mathbb{R}^{n \times n}$ (matrix of pairwise distances), $p \in [n]$ (reduced dimension).

$V \leftarrow$ matrix from $\mathbb{R}^{n \times n}$ with $2(n-1)$ on the diagonal and -2 elsewhere

$V^+ \leftarrow$ Moore-Penrose inverse of V

$\mathbf{Z} \leftarrow$ random matrix from $\mathbb{R}^{n \times p}$ *(initialization)*

while not converged **do**

$\{z_i\}_{1 \leq i \leq n} \leftarrow$ rows of $\mathbf{Z}$

$\delta_{ij} \leftarrow \|z_i - z_j\|_{\ell_2}$ for all $(i,j) \in [n]$

$V' \leftarrow$ matrix from $\mathbb{R}^{n \times n}$ with $\left(2 \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0}\right)_{1 \leq i \leq n}$ on the diagonal and $\left(-2 \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0}\right)_{1 \leq i \neq j \leq n}$ elsewhere.

$\mathbf{Z} \leftarrow V^+ V' \mathbf{Z}$

end while

$\{z_i\}_{1 \leq i \leq n} \leftarrow$ rows of $\mathbf{Z}$

Output: $\{z_i\}_{1 \leq i \leq n}$.