Recitation Week 14

Ashwin Bhola

CDS, NYU

Dec 4th , 2019

Gradient Descent

1. Given N training points $\{(x_k, y_k)\}_{k=1}^{N}$, $x_k \in \mathbb{R}^d$, $y_k \in \{-1,1\}$, we seek a linear discriminant function $f(x) = w^T x$

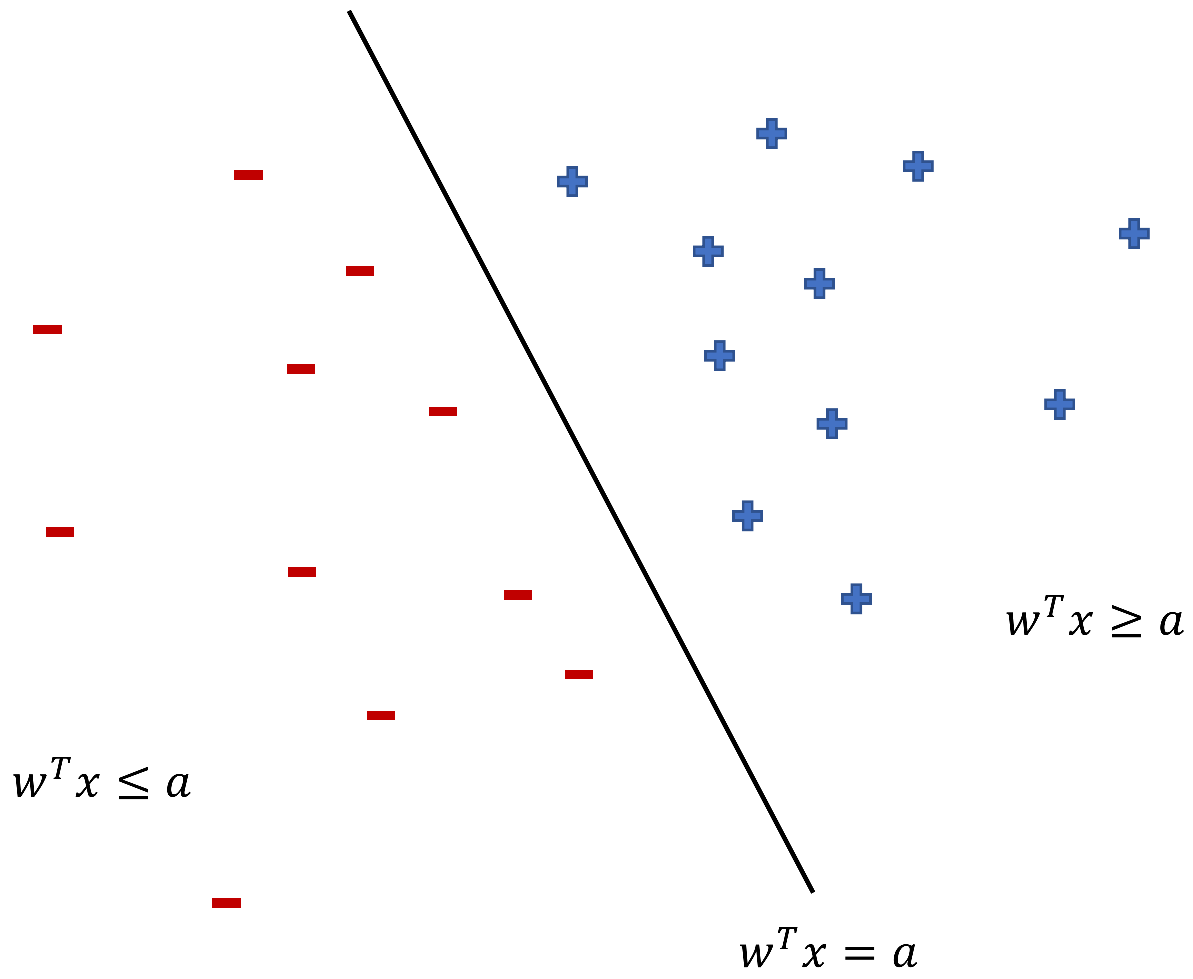
   i. Suggest a method for determining if the points are perfectly linearly separable

   ii. Let's say we decide to take the loss function as the exponential loss, $L(y, f(x)) = e^{-z}$, where z is defined to be the margin, $z = yf(x)$, y being the label of x and $f(x)$ being our prediction on x. Derive the update rule for batch gradient descent for exponential loss. What's the computational cost of this update in terms of N and d?

   iii. The step size in gradient descent is not actual "step size" because the true length of the step in gradient descent will be $\alpha||\nabla_w J(w)||$. Someone suggests that we should scale the gradients to make them unit norm before we do gradient descent. What's the problem with this?

   iv. What are the different possible stopping criteria for gradient descent algorithm?

Gradient Descent



Gradient Descent

i. If there exists $w \in \mathbb{R}^d$ such that $\mathrm{y_i} w^T x_i > 0 \; \forall \, i$, then the points are linearly separable

ii. $J(w) = \Sigma_i L\big(y_i, f(x_i)\big) = \Sigma_i e^{-y_i w^T x_i}$

$$w_{t+1} = w_t - \alpha \nabla J(w)$$
$$\nabla J(w) = \sum_i (-y_i x_i) e^{-y_i w^T x_i}$$
$$\Rightarrow w_{t+1} = w_t + \alpha \sum_i y_i x_i e^{-y_i w^T x_i}$$

iii. If you normalize the gradient every time, you'll lose the advantage of decreasing step size unless you schedule α to also decrease as you approach the minimum

iv. You can continue doing gradient for predetermined number of iterations or choose one of these: $\left|\left|w_{t+1} - w_t\right|\right| < \epsilon, |J(w_{t+1}) - J(w_t)| < \epsilon, \left|\left|\nabla J(w)\right|\right| < \epsilon$. $|J(w_{t+1}) - J(w_t)| < \epsilon$ is preferred

Optimality conditions

1. Consider f to be twice differentiable function. True or False:

   i. For a convex function, the first order Taylor approximation at any point is a global under estimator of the function

   ii. Convex functions do not have saddle points

   iii. For x to be a local minimum of a function $f(x)$, $\nabla f(x) = 0 \; and \; \mathrm{H}_f(x) \succ 0$

   iv. The necessary and sufficient condition for local optimality in unconstrained convex optimization is $\nabla f(x) = 0$

   v. The necessary and sufficient condition for local optimality in unconstrained optimization is $\nabla f(x) = 0$

   vi. For convex functions, the direction of steepest descent is same as the direction towards global optima

   $vii. \; e^y \geq 1 + y$

   viii. Standardizing data does not help gradient descent

Optimality conditions

1. Consider f to be twice differentiable function. True or False:

   i. True: $f(y) \geq f(x) + \nabla f(x)^T (y - x)$

   ii. True: $If\ \nabla f(x) = 0$, and f is a convex function, then x is the minimizer

   iii. False: x^4

   iv. True

   v. False: You need second order conditions

   vi. False

   vii. True: Use part i

   viii. False: Standardizing brings all eigenvalues of covariance matrix to the same scale