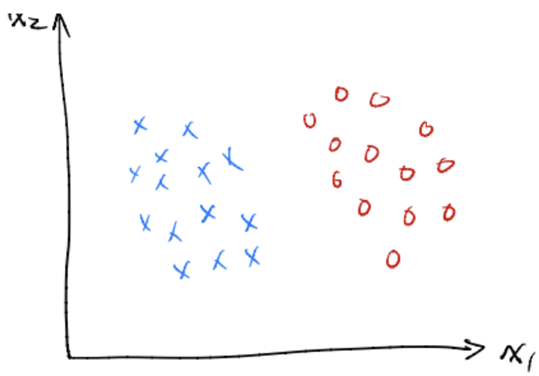
Lecture 3 SVM

wbg231

December 2022

1 maximum margin classifier



-
- consider trying to learn a classifier for these linearly separable classes
- that task can be written as find a vector $w \in \mathbb{R}^d : \forall i \in [1, n](w^t x_i)y_i > 0$
- keep in mind that a hyperplane is defined as the sec of vectors orthogonal to our vector w (plus some offset) that is $\{v \in \mathbb{R}^d : w^t v + b = 0\}$
- we can do this simply with the Perceptron algorithm

  - Initialize $w \leftarrow 0$
  - While not converged (exists misclassified examples)
    - For $(x_i, y_i) \in \mathcal{D}$
      - If $y_i w^T x_i < 0$ (wrong prediction)
      - Update $w \leftarrow w + y_i x_i$

- this more or less moves our hyperplane towards the misclassified points, as long as the data is linearly separable this will work

- geometric margin given a hyperplane G that separates a dataset the geometric margin is the distance to the hyperplane of the closest data point in our dataset that is
$$min_i d(x_i, H)$$
- we know the vector w is orthogonal to the hyperplane so any point projected on w will be orthogonal to the hyperplane that. so that is for an arbitrary vector v
$$\frac{< v, w >}{||w||_2}$$
- we know that this projection is orthogonal to H and thus parallel to w ie $\frac{<v,w>}{||w||_2} = \lambda w$
- notice that $x - h = P_{h^\perp}(x) = P_w(h) = \lambda w$
- meaning we can write $d(x, h) = ||x - h||_2 = ||\lambda w|| = |\lambda|||w||_w = |\frac{w^t x+b}{||w||_2}|$ b here is an offset term.

minimize the margin

- our goal is to max the geometric margin that if
$$max(min_i(d(x_i, H))) = maxmin_i \frac{y_i(w^t x_i + b)}{||w||_2}$$
- can re write as constrained max as
$$maxM$$
$$\text{subject to } \frac{y_i(w^t x_i + b)}{||w||_2} \leq M \forall i$$
- note here that this will not give us a unique solution since this is not scale invariant
- we can force uniqueness by adding a constraint to the norm of w that is let $||w||_2 = \frac{1}{M}$ which allows us to write
$$max \quad \frac{1}{||w||_2}$$
$$\text{subject to } y_i(w^t x_i + b) \geq 1$$
$$\iff$$
$$min \quad \frac{1}{2}||w||_2^2$$
$$\text{subject to } y_i(w^t x_i + b) \geq 1$$

- find max norm solution such that the functional margin is greater than 1 for all examples
- we can make this a soft margin classifier by adding a slack term which makes the problem

$$min \quad \frac{1}{2}||w||_2^2 + \frac{C}{n}\sum_{i=1}^{n}\epsilon_i$$

$$\text{subject to } y_i(w^t x_i + b) \geq 1 - \epsilon_i \forall u, \epsilon_i \geq 0 \forall i$$

- ϵ are the slack we are given each exmample (ie how much we relax the margin constraint for it)
- C is a weighing term that penalizes more $||\epsilon_i||_1$

minimize hinge loss

- Perceptron loss is $\ell(x, y, w) = max(0, -yw^t x)$ that is zero if all points are correctly classified
- hinge loss is $\ell_{hinge}(x, w, y) = max(0, 1 - m) = max(0, 1 - y(w^t x))$ so we linearly penalize solution until they achieve a functional margin of 1 then ignore them
- we can write SVM problem in terms of ERM over a linear hypotheses space plus and offset term (can also think of the space as the set of hyperplanes)
- with hinge loss
- and l2 regularization
- that is

$$j(w, b) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(w, b, x)) + \lambda||w||_2 = \frac{c}{n}max(0, 1 - y_i(w^t x_i + b)) + \frac{1}{2}||W||_2^2$$

- can clearly re-write this as a constrained optimization problem as

$$min \quad \frac{1}{2}||w||_w^2 + \frac{c}{n}\sum_{i=1}^{n}\epsilon_i$$

$$\text{subject to } \epsilon_i \geq max(0, 1 - y_i(w^t x_i + b))$$

- so we can derive the objetive in either way and our new problem is to optimize it

sub-gradient descent

- subgradient more or less generalizes a taylor expansion
- a vector $g \in \mathbb{R}^d$ is a subgradient of a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ at x if for all z
$$f(z) \geq f(x) + g^t(z - x)$$
- so it just means the vector ie straight line implied by that function is bellow the function for all values of x

- Initialize $w \leftarrow 0$
- While not converged (exists misclassified examples)
  - For $(x_i, y_i) \in \mathcal{D}$
    - If $y_i w^T x_i < 0$ (wrong prediction)
    - Update $w \leftarrow w + y_i x_i$

- subgradient descent move along a negative subgradient g that is
$$x^{t+1} = x^t - \eta g \text{ where } g \in \partial f(x^t) \text{ and } \eta > 0$$
- this can increase the objective function but will always get us closer to the arg min if f is convex and step size is small
- it is slower than gradient descent but sometimes our best option
- so given our svm objective
$$j(w, b) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(w, b, x)) + \lambda||w||_2 = \frac{c}{n} max(0, 1 - y_i(w^t x_i + b)) + \frac{1}{2}||W||_2^2$$
- we can see that the a subgradient is given by $2\lambda w$ if $1 - y_i w^t x_i \leq 1$ and $2\lambda w + y_i x_i$ other wise
- thus our subgradient descent algorithm is

- Initialize $w \leftarrow 0$
- While not converged (exists misclassified examples)
  - For $(x_i, y_i) \in \mathcal{D}$
    - If $y_i w^T x_i < 0$ (wrong prediction)
    - Update $w \leftarrow w + y_i x_i$

dual problem

- recall that in a geneal optimization problem with inequality constraints can be expressed as
$$min \quad f_0(x)$$
$$\text{subject to } f_i(x) \leq 0 \forall i$$

-

- we can write the Lagrangian form as
$$\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i()x$$
this is a weighted sum of the objective and constraint functions, this process can treat hard constraints as soft ones

- this defines the Lagrange dual function as
$$g(\lambda) = inf_x \mathcal{L}(x, \lambda) = inf_x(f_0(x) + \sum i = 1^m \lambda_i f_i(x))$$
this has some nice properties namely that it is concave, and is a lower bound for our optimization problem

- weak duality tells us that it the dual solution si a lower bound of the primal

- strong duality says the solutions are equal

- if we have strong duality then we knw $f_0(x^*) = G(\lambda^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \leq f_0(x^*)$

- in other words $\sum_i \lambda^* f_i(x^*) = 0$ meaning that
$$\lambda_i > 0 \Rightarrow f_i(x^*) = 0 \text{ and } f_i(x^*) < 0 \Rightarrow \lambda_0 = 0$$

- so our constraints are only active when our objective is zero, and our constraints are inactive when our objective is less than zero

- the SVM primal
$$min \quad \frac{1}{2}||w||_w^2 + \frac{c}{n} \sum_{i=1}^{n} \epsilon_i$$
$$\text{subject to } -\epsilon_i \leq -\forall$$
$$(1 - y_i[w^t x_i + b]) - \epsilon_i \leq 0 \forall i$$

- svm has strong duality since the problem is convex (as long as we have feasible points)

- the svm dual is

$$sup_{\alpha\geq 0,\lambda\geq 0} inf_{w,b,\epsilon} L(w,\underline{\ },\epsilon,\alpha,\lambda)$$

$$= sup_{\alpha\geq 0,\lambda\geq 0} inf_{w,b,\epsilon} \frac{1}{2}||w||_2^2 + \frac{c}{n}\sum_{i=1}^{n}\epsilon_i + \sum_{i=1}^{n}\alpha_i(1-y_i(w^t x_i+b)-\epsilon_i) + \sum_{i=1}^{n}\lambda_i(-\epsilon_i)$$

- solving this out yields that $\epsilon i = max(0, 1 - y_i f*(x_i))$ is the hingle loss on that training example
- it $y_i f^*(x_i) > 1$ then the margin loss is $\epsilon_i = 0$ and we get $\alpha_i = 0$ (ie points we can correctly classified are fine)
- if $y(f^*(x_i)) < 1$ then the margin loss is $\epsilon_i > -$ and $a\alpha_i = \frac{c}{n}$ (that makes sense we are adding a weighting constant)
- if $\alpha_i = 0$ that is we know there is no loss ie $y_i f^*(x) \geq 1$
- if $\alpha_i \in (0, \frac{c}{n})$ then $\epsilon_i = 0$ meaning that our point is on the margin ie $1 - y_i f*(x_i) = 0$
- support vectors are training points such that $\alpha_i \in [0, \frac{c}{n}]$ that is are on the marign
- if there are few margin ie few support vectors then we will have sparsity in input examples as $w^* = \sum_{i=1}^{n}\alpha_i^* y_i x_i$
- so if $\alpha_i = 0$ we dont weight the example
- if $\alpha_i = \frac{c}{n}$ then $y_i f(x_i) \leq 1$ so we weigh it
- if $y_i f(x_i) < 1$ then $\alpha_i = \frac{c}{n}$