MATERIAL FOR SUPPORT VECTOR MACHINES
MACHINE LEARNING E20

Mathias Ravn Tversted
Department of Computer Science
Aarhus University
Åbogade 34, 8200 Aarhus N
Tversted@post.au.dk

December 27, 2020

Contents

1 Intro and motivation

Suppose we have some linear model, such as in binary classification with $x = (1, x_1, \cdots, x_d)$ with labels $y \in \{-1, 1\}$. Where prediction is $w(x) = \text{sign}(w^T x)$. The PLA is good for classifying training data, but not all seperating lines are equally good when trying to generalize for new points. Here we wish to select the line that has the largest margin to the two groups.

TEGN EKSEMPEL PÅ AT NOGLE STREGER ER BEDRE END ANDRÈ

2 Margins

First we describe a *functional* margin, which is the distance to the hyperplane.

2.1 Functional margins

:

- Points are (x_i, y_i)
- $\hat{y_i} = y_i(w^T x_i + b)$
- $\hat{y} = min_i \hat{y_i}$

Remember that w is the vector that is orthogonal to the seperating hyperplane. We wish to minimize this margin. A problem with this definition of the margin is that multiple values of w and b describe the same hyperplane, such as $w^T x + b = 0 = (nw)^T x + nb$, potentially giving you arbitrarily large weights and biases if you attempt to maximise the margin.

2.2 Geometric margins

Instead we formalize a margin that is invariant to scale.
$x - \delta w$ is the projection of x onto the hyperplane. (We use that w is orthogonal on the hyperplane).
$w^T(x - \delta w) + b$ must be equal to 0 since the distance to the hyperplace is 0.

$$\begin{aligned}
w^T(x - \delta w) + b &= 0 && (1)\\
w^T x - \delta \|w\|^2 + b &= 0 && \text{(Multiply into parenthesis)} \quad (2)\\
w^T x / \|w\| - \delta \|w\| + b/\|w\| &= 0 && \text{(Scale down by } \|w\|\text{)} \quad (3)\\
w^T x / \|w\| + b/\|w\| &= \delta \|w\| && \text{(Add } \delta\|w\| \text{ to both sides)} \quad (4)\\
& && (5)
\end{aligned}$$

Now we have an equation for the signed distance to the hyperplane which doesn't have the problem with multiple (w, b) describing the same hyperplane, because orthogonal projections are unique if they exist.

3 Linear System

We wish to solve the following linear system to find the hyperplane that maximizes the margin while staying scale invariant.

$$\begin{aligned}
&\max y && && (6)\\
&\text{s.t } y_i(w^T x_i + b) \geq y && i = 1, \dots, n && (7)\\
&\|w\| = 1 && && (8)
\end{aligned}$$

The same optimization problem can be expressed as a minimisation problem that minimizes the norm of w.

$$\min \frac{1}{2}\|w\|^2 \tag{9}$$

$$\text{s.t } y_i(w^T x_i + b) \geq 1 \qquad i = 1, \ldots, n \tag{10}$$

MANGLER ARGUMENT FOR HVORFOR DE ER DE SAMME

4 Note: About primal and dual

If our objective function f is convex and all of our g_i constraints are strictly feasible (x where $g_i(x) < 0 \forall i$) then $d^* = p^*$

Karush-Kuhn-Tucker conditions: There is an optimal solution x^*, a^* satisfying

1. $\frac{\partial L(x*,a*)}{\partial x_i} = 0, \text{ for } i = 1, \cdots, d$
2. $a_i^* g_i(x^*) = 0, \text{ for } i = 1, \cdots, n$
3. $g_i(x^*) \leq 0, \text{ for } i = 1, \cdots, n$
4. $a_i^* \geq 0, \text{ for } i = 1, \cdots, n$

5 Rewriting problem - Dette afsnit er cursed og skal sikkert gennemgås igen

The form of a general optimisation problem is

$$\min_x f(x) \tag{11}$$

$$\text{s.t } g_i(x) \leq 0 \qquad i = 1, \cdots, n \tag{12}$$

And so our SVM in standard form becomes

$$\min_x \frac{1}{2}\|w\|^2 \tag{13}$$

$$\text{s.t } g_i(w, b) \leq 0 \qquad \text{with} \tag{14}$$

$$g_i(w, b) = 1 - y_i(w^T x_i + b) \tag{15}$$

if f and g_i are convex, we can solve dual and assume the solutions satisfy *KKT* conditions.

They are! (Proofs omitted). We don't need to proof this, but if we were just look at the hessian matrix. It's all 0 uwu.

Furthermore, all of our constraints need to be strictly feasible, but if the data is linearly seperable, then there exists a hyperplane which correctly seperates all of the points, then clearly

$$y_i(w^T x_i + b) > 0 \qquad \forall i \tag{16}$$

And $g_i(w, b) < 0$ if w, b are scaled by appropriate choice of constant c.

We will thus solve the dual

$$\max_{a:a_i \geq 0} \min_{w,b} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a(y_i(w^T x_i + b) - 1) \tag{17}$$

Final formulation: We want to recover the optimal w and b, and we can achieve this by setting derivatives to 0 as in the *KKT* conditions (the first), and likewise the optimal b.

The w we get is a hyperplane that satisfied $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ and the optimal b is halfway from the closest blue and furthest red in w's direction. $b = -\frac{1}{2}(\min_{i:y_i=1} \frac{w^T x_i}{\|W\|} + \max_{i:y_i=-1} \frac{w^T x_i}{\|w\|})$

The final optimisation problem is thus

$$\max_{a:a_i \geq 0} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{j=1}^{n} a_i a_j y_i y_j x_i^T x_j \tag{18}$$

$$\text{s.t} \sum_{i=1}^{n} a_i y_i = 0 \tag{19}$$

Here values a_i is non-zero when $y_i(w^T x_i + b) = 1$, which means (x_i, y_i) is on the margin. w is a linear combination of only the support vectors.

6 Kernels

Classifcation with sign$(w^T x + b)$ takes $O(d)$ time, but we can also do

$$\text{sign}((\sum_{a_i \neq 0} a_i y_i x_i)^T x + b) = \text{sign}(\sum_{a_i \neq 0} a_i y_i x_i^T x + b) \tag{20}$$

Which takes $O(kd)$ for k support vectors.

While this seems slower, it may be an advantage if we have to rely on non-linear feature transforms.

If we have some transform $\phi : R^d \to R^{d'}$, we can create some kernel which turns $\phi(X_i)^T \phi(x)$ into some Kernel$(x, y) : R^d \times R^d \times \to R$, where the kernel computes the inner product of the inner products.

We can also reformulate our problem *again* so that we find the maximum seperating margin in *feature space*, which we will obviously need to do if the data is not linearly seperable.

$$\max_{a:a_i \geq 0} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{j=1}^{n} a_i a_j y_i y_j \text{kernel}(x_i, x_j) \tag{21}$$

$$\text{s.t} \sum_{i=1}^{n} a_i y_i = 0 \tag{22}$$

With bias $b = -\frac{1}{2}(\min_{i:y_i=1} \frac{w^T x_i}{\|w\|} + \max_{i:y_i=-1} \frac{w^T x_i}{\|w\|})$