TTIC 31250
An Introduction to the Theory of Machine Learning

Avrim Blum

04/15/20

Lecture 4: Support Vector Machines

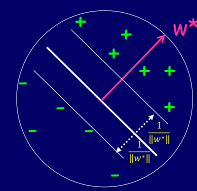
1

Perceptron Recap

Perceptron alg makes at most $\|w^*\|^2 R^2$ mistakes if $\exists w^*$ with $w^* \cdot x \geq 1$ on all positives and $w^* \cdot x \leq -1$ on all negatives, and all $\|x\| \leq R$.

Algorithm:

- Initialize $w = \vec{0}$. Predict positive if $w \cdot x > 0$, else predict negative.
- Mistake on positive: $w \leftarrow w + x$.
- Mistake on negative: $w \leftarrow w - x$.

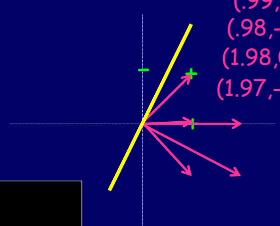


2

Perceptron algorithm

Example: (0.01,1) –
(1,1) +
(1,0) +

(1,1)
(.99,0)
(.98,-1)
(1.98,0)
(1.97,-1)



Algorithm:

Initialize $w = \vec{0}$. Use w · x > 0.

- Mistake on pos: w ← w+x.
- Mistake on neg: w ← w-x.

3

Analysis

Perceptron alg makes at most $\|w^*\|^2 R^2$ mistakes if $\exists w^*$ with $w^* \cdot x \geq 1$ on all positives and $w^* \cdot x \leq -1$ on all negatives, and all $\|x\| \leq R$.

Proof: consider w · w* and ||w||

- Each mistake increases w · w* by at least 1.
  (w + x) · w* = w · w* + x · w* ≥ w · w* + 1.
  So after M mistakes, $w \cdot w^* \geq M$.
- Each mistake increases w·w by at most R^2.
  (w + x) · (w + x) = w·w + 2(w·x) + x·x ≤ w·w + R^2.
  So, after M mistakes, $\|w\|^2 \leq MR^2$, so $\|w\| \leq \sqrt{M}R$.

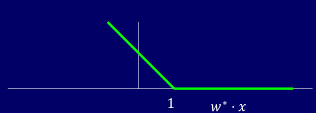
Since $\frac{w \cdot w^*}{\|w^*\|} \leq \|w\|$, get $\frac{M}{\|w^*\|} \leq \sqrt{M}R$ so $\sqrt{M} \leq \|w^*\|R$.

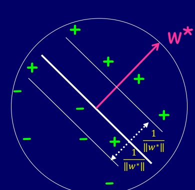
4

What if w* isn't perfect?

Theorem: on any sequence of examples S, the Perceptron algo makes at most $\min_{w^*}[\|w^*\|^2 R^2 + 2L_{hinge}(w^*, S)]$ mistakes.

The hinge-loss of w* on x is the amount by which the desired inequality ($w^* \cdot x \geq 1$ or $w^* \cdot x \leq -1$) is not satisfied.





5

Support Vector Machines (SVMs)

In the batch (PAC) setting, we are given S up front. Let's just solve for w^* of largest margin. ("realizable case")

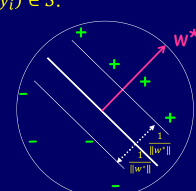
Convex optimization problem:

Minimize: $\|w\|^2$

Subject to: $y_i(w \cdot x_i) \geq 1$ for all $(x_i, y_i) \in S$.

(viewing y_i as ± 1)

But what if there's no perfect separator?



6

Support Vector Machines (SVMs)

Let's solve for the solution that minimizes a (generalization of) the Perceptron mistake bound.

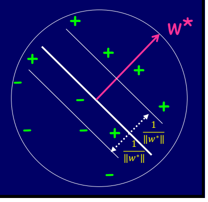
Given a quantity C as input:

Minimize: $\|w\|^2 + C \boxed{\sum_{x_i \in S} \xi_i}$ hinge loss. The ξ_i are "slack variables"

Subject to: $y_i(w \cdot x_i) \geq 1 - \xi_i$ for all $(x_i, y_i) \in S$.

$\xi_i \geq 0$ for all i.

This is the SVM algorithm. The quantity C trades off margin and hinge-loss.



7

Support Vector Machines (SVMs)

Given a quantity C as input:

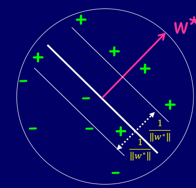
Minimize: $\|w\|^2 + C \boxed{\sum_{x_i \in S} \xi_i}$ hinge loss. The ξ_i are "slack variables"

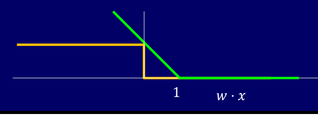
Subject to: $y_i(w \cdot x_i) \geq 1 - \xi_i$ for all $(x_i, y_i) \in S$.

$\xi_i \geq 0$ for all i.

Some intuition:

- The total hinge loss is an upper bound on empirical 0/1-loss (# mistakes on S) of the classifier $w \cdot x > 0$.





8

Support Vector Machines (SVMs)

Given a quantity C as input:

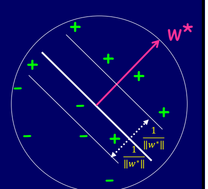
Minimize: $\|w\|^2 + C \boxed{\sum_{x_i \in S} \xi_i}$ hinge loss. The ξ_i are "slack variables"

Subject to: $y_i(w \cdot x_i) \geq 1 - \xi_i$ for all $(x_i, y_i) \in S$.

$\xi_i \geq 0$ for all i.

Some intuition:

- The total hinge loss is an upper bound on empirical 0/1-loss (# mistakes on S) of the classifier $w \cdot x > 0$.
- The first term $\times R^2$ is roughly (take on faith for now) an upper bound on the amount of overfitting.
- Together, proportional to rough upper-bound on true error.



9

Support Vector Machines (SVMs)

Given a quantity C as input:

Minimize: $\|w\|^2 + C \boxed{\sum_{x_i \in S} \xi_i}$ hinge loss. The ξ_i are "slack variables"

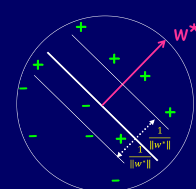
Subject to: $y_i(w \cdot x_i) \geq 1 - \xi_i$ for all $(x_i, y_i) \in S$.

$\xi_i \geq 0$ for all i.

This is the *primal* form of SVM.

To kernelize it, we will want to move to the *dual* form.

So, first a bit about the Lagrangian dual…



10

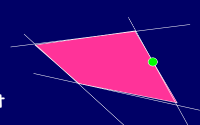
Lagrangian Dual

Consider an optimization problem of the form:

Minimize: convex function in some variables (like w_i, ξ_i)

Subject to: linear constraints on these variables.

Think of as a game between a corporation that wants to minimize its costs (given by the convex function) and a government, that doesn't want the corporation to break any laws (given by the linear constraints).



11

Lagrangian Dual

Consider an optimization problem of the form:

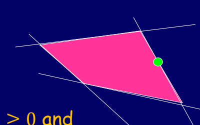
Minimize: convex function in some variables (like w_i, ξ_i)

Subject to: linear constraints on these variables.

For each constraint, the govt can charge a fine that is linear in the amount by which it is violated.

E.g., if govt puts fine of \$100 on $\xi_i \geq 0$ and corp uses $\xi_i = -0.5$ then corp pays \$50.

But there's a catch: must be fully linear. If corp uses $\xi_i = +0.5$ then corp collects \$50.



12

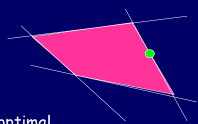
Lagrangian Dual

Consider an optimization problem of the form:

Minimize: convex function in some variables (like w_i, ξ_i)

Subject to: linear constraints on these variables.

The game: for each constraint j, govt gets to choose $\alpha_j \geq 0$ linear penalty. Corp chooses setting of variables. Corp wants to minimize cost. Govt wants to maximize (or equivalently to keep corp honest).



If corp has to go first, clearly should pick optimal feasible point (else govt will assign infinite penalty to any violated constraint).

Claim: If govt goes first, can assign penalties such that corp can do no better. (No "duality gap".) This relies on convexity of the cost function. Govt's optimization problem is called the dual.

13

Lagrangian Dual

Let $\vec{w}$ denote strategy of corp (primal variables) and let $\vec{\alpha}$ denote strategy of govt (dual variables). The Lagrangian is the total cost $L(\vec{w}, \vec{\alpha})$ paid by corp.

Govt's optimization problem is:

$$\max_{\vec{\alpha}} \min_{\vec{w}} L(\vec{w}, \vec{\alpha}) \quad \textit{subject to} \quad \alpha_j \geq 0 \ \ \forall j$$

Let's see how this plays out for SVMs.

14

SVM Dual Formulation

Primal: Minimize: $\frac{1}{2}\|w\|^2 + C\ \sum_i \xi_i$

Subject to: $y_i(w \cdot x_i) \geq 1 - \xi_i$ for all $(x_i, y_i) \in S$.

$\xi_i \geq 0$ for all i.

Lagrangian: have variables $\alpha_{i1} \geq 0$ for each constraint of 1st kind and $\alpha_{i2} \geq 0$ for each constraint of 2nd kind.

Govt's optimization problem is:

$$\max_{\alpha_1, \alpha_2} \min_{w, \xi} \frac{1}{2}\|w\|^2 + C \sum_i \xi_i + \sum_i \alpha_{i1}\big(1 - \xi_i - y_i(w \cdot x_i)\big) - \sum_i \alpha_{i2}\xi_i$$

subject to $\alpha_{i1}, \alpha_{i2} \geq 0$ for all i.

15

SVM Dual Formulation

Now, let's think about a specific ξ_i. Contribution is $\xi_i(C - \alpha_{i1} - \alpha_{i2})$. Govt had better set $\alpha_{i1} + \alpha_{i2} = C$, else corp can make this $-\infty$. So, replace α_{i2} with $C - \alpha_{i1}$, let $\alpha_i = \alpha_{i1}$, and have constraint $0 \leq \alpha_i \leq C$. Simplifies to…

Govt wants to solve for α_i s.t. $0 \leq \alpha_i \leq C$ to maximize

$$\min_w \frac{1}{2}\|w\|^2 + \sum_i \alpha_i(1 - y_i(w \cdot x_i))$$

Govt's optimization problem is:

$$\max_{\alpha_1, \alpha_2} \min_{w, \xi} \frac{1}{2}\|w\|^2 + C \sum_i \xi_i + \sum_i \alpha_{i1}\big(1 - \xi_i - y_i(w \cdot x_i)\big) - \sum_i \alpha_{i2}\xi_i$$

subject to $\alpha_{i1}, \alpha_{i2} \geq 0$ for all i.

16

SVM Dual Formulation

Now, let's think about a specific ξ_i. Contribution is $\xi_i(C - \alpha_{i1} - \alpha_{i2})$. Govt had better set $\alpha_{i1} + \alpha_{i2} = C$, else corp can make this $-\infty$. So, replace α_{i2} with $C - \alpha_{i1}$, let $\alpha_i = \alpha_{i1}$, and have constraint $0 \leq \alpha_i \leq C$. Simplifies to…

Govt wants to solve for α_i s.t. $0 \leq \alpha_i \leq C$ to maximize

$$\min_w \frac{1}{2}\|w\|^2 + \sum_i \alpha_i(1 - y_i(w \cdot x_i))$$

We can solve inner minimization by setting gradient to 0:

$$w - \sum_i \alpha_i y_i x_i = 0.$$

Plug in $w = \sum_i \alpha_i y_i x_i$ above.

17

SVM Dual Formulation

Dual: solve for α_i s.t. $0 \leq \alpha_i \leq C$ to maximize

$$\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_i \alpha_i - \sum_i \alpha_i y_i \sum_j \alpha_j y_j (x_j \cdot x_i).$$

$$= \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).$$

Govt wants to solve for α_i s.t. $0 \leq \alpha_i \leq C$ to maximize

$$\min_w \frac{1}{2}\|w\|^2 + \sum_i \alpha_i(1 - y_i(w \cdot x_i))$$

We can solve inner minimization by setting gradient to 0:

$$w - \sum_i \alpha_i y_i x_i = 0.$$

Plug in $w = \sum_i \alpha_i y_i x_i$ above.

18

SVM Dual Formulation

Dual: solve for α_i s.t. $0 \le \alpha_i \le C$ to maximize

$$\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_i \alpha_i - \sum_i \alpha_i y_i \sum_j \alpha_j y_j (x_j \cdot x_i).$$

$$= \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i \cdot x_j).$$

Notice this is kernelizable. Hence, we can run SVMs with any kernel using the dual and replacing $x_i \cdot x_j$ with $K(x_i, x_j)$.

19

Intro to Tail Inequalities

20

Chernoff and Hoeffding bounds

Consider m flips of a coin of bias p. Let N_{heads} be the observed # heads. Let $\varepsilon, \alpha \in [0,1]$.

Hoeffding bounds:

- $\Pr[N_{heads}/m > p + \varepsilon] \le e^{-2m\varepsilon^2}$
- $\Pr[N_{heads}/m < p - \varepsilon] \le e^{-2m\varepsilon^2}$

Chernoff bounds:

- $\Pr[N_{heads}/m > p(1+\alpha)] \le e^{-mp\alpha^2/3}$
- $\Pr[N_{heads}/m < p(1-\alpha)] \le e^{-mp\alpha^2/2}$

E.g,

- $\Pr[N_{heads} > 2(\text{expectation})] \le e^{-(\text{expectation})/3}$.
- $\Pr[N_{heads} < (\text{expectation})/2] \le e^{-(\text{expectation})/8}$.

21

Typical use of bounds

Thm: If $|S| \ge \frac{1}{2\epsilon^2}\left[\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)\right]$, then with prob $\ge 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

- Proof: Just apply Hoeffding + union bound.
  - Chance of failure at most $2|H|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.

Hoeffding bounds:

- $\Pr[N_{heads}/m > p + \varepsilon] \le e^{-2m\varepsilon^2}$
- $\Pr[N_{heads}/m < p - \varepsilon] \le e^{-2m\varepsilon^2}$

22

Typical use of bounds

Thm: If $|S| \ge \frac{1}{2\epsilon^2}\left[\ln(2|H|) + \ln\left(\frac{1}{\delta}\right)\right]$, then with prob $\ge 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

- Proof: Just apply Hoeffding + union bound.
  - Chance of failure at most $2|H|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.
- So, whp, best on sample is ε-best over D.
  - Note: this is worse than previous bound ($1/\varepsilon$ has become $1/\varepsilon^2$), because we are asking for something stronger.
  - Can also get bounds "between" these two.

23

Typical use of bounds

Thm: If $|S| \ge \frac{6}{\epsilon}\left[\ln|H| + \ln\frac{1}{\delta}\right]$, then with prob $\ge 1-\delta$, all $h \in H$ with $err_D(h) > 2\varepsilon$ have $err_S(h) > \varepsilon$, and all $h \in H$ with $err_D(h) < \varepsilon/2$ have $err_S(h) < \varepsilon$.

Proof: apply Chernoff…

Chernoff bounds:

- $\Pr[N_{heads}/m > p(1+\alpha)] \le e^{-mp\alpha^2/3}$
- $\Pr[N_{heads}/m < p(1-\alpha)] \le e^{-mp\alpha^2/2}$

E.g,

- $\Pr[N_{heads} > 2(\text{expectation})] \le e^{-(\text{expectation})/3}$.
- $\Pr[N_{heads} < (\text{expectation})/2] \le e^{-(\text{expectation})/8}$.

24