Lecture 4: Support Vector Machines

Perceptron Recap

Perceptron algorithm 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 \( w = 0 \). Predict positive if \( w \cdot x > 0 \), else predict negative.
- Mistake on positive: \( w \leftarrow w + x \).
- Mistake on negative: \( w \leftarrow w - x \).

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

Since \( \frac{w^*}{\|w^*\|} \leq \|w\| \), get \( \frac{M}{\|w^*\|} \leq \sqrt{MR} \) so \( \sqrt{M} \leq \|w^*\| R \).

What if \( w^* \) isn’t perfect?

Theorem: on any sequence of examples \( S \), the Perceptron algo makes at most \( \min \{ \|w^*\|^2 R^2 + 2 \text{ hinge}(w^*, S) \} \) 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.

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 \( \pm 1 \))

But what if there’s no perfect separator?
**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: $||w||^2 + C \sum_{i \in S} \xi_i$

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 $i$.

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

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 $||w||^2$ is roughly (take on faith for now) an upper bound on the amount of overfitting.
- Together, proportional to rough upper-bound on true error.

**Lagrangian Dual**

Consider an optimization problem of the form:

Minimize: convex function in some variables (like $w, \xi_i$)

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).

Consider an optimization problem of the form:

Minimize: convex function in some variables (like $w, \xi_i$)

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 $50$. 

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.
Consider an optimization problem of the form:

Minimize: convex function in some variables (like $w$, $\xi$)
Subject to: linear constraints on these variables.

The game: for each constraint, govt gets to choose $\alpha_i \geq 0$ 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.

Let's see how this plays out for SVMs.

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

We can solve inner minimization by setting gradient to 0:

$$w = \sum \alpha_i y_i x_i$$

Plug in $w = \sum \alpha_i y_i x_i$ above.
SVM Dual Formulation

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

$$\frac{1}{2} \sum_{i,j} a_i a_j y_i y_j (x_i \cdot x_j) + \sum_i a_i - \sum_j a_j y_j (x_j \cdot x_i).$$

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

Intro to Tail Inequalities

Chernoff and Hoeffding bounds

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

Hoeffding bounds:
- $\Pr[N_{\text{heads}} / m > p + \sigma] \leq e^{-2m\sigma^2}$
- $\Pr[N_{\text{heads}} / m < p - \sigma] \leq e^{-2m\sigma^2}$

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

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

Typical use of bounds

Thm: If $|S| \geq \frac{1}{2\epsilon^2} \left[ \ln(2|S|) + \ln \left( \frac{2}{\epsilon} \right) \right]$, then with prob $\geq 1 - \delta$, all $h \in H$ have $|\text{err}_0(h) - \text{err}_D(h)| < \epsilon$.

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

Hoeffding bounds:
- $\Pr[N_{\text{heads}} / m > p + \sigma] \leq e^{-2m\sigma^2}$
- $\Pr[N_{\text{heads}} / m < p - \sigma] \leq e^{-2m\sigma^2}$

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

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

Typical use of bounds

Thm: If $|S| \geq \frac{1}{2\epsilon^2} \left[ \ln(2|S|) + \ln \left( \frac{2}{\epsilon} \right) \right]$, then with prob $\geq 1 - \delta$, all $h \in H$ with $\text{err}_0(h) > 2\epsilon$ have $\text{err}_D(h) > \epsilon$.

Proof: apply Chernoff...