Support vector machines (SVMs)
Lecture 4

David Sontag
New York University

Slides adapted from Luke Zettlemoyer, Vibhav Gogate, and Carlos Guestrin
Key idea #1: Allow for slack

\[ w \cdot x + b = +1 \]
\[ w \cdot x + b = 0 \]
\[ w \cdot x + b = -1 \]

\[ \sum_j \xi_j \geq 0 \]

"slack variables"

Solve for the optimal value \( \xi_j^* \) as a function of \( w \) and \( b \):

If \((w \cdot x_j + b) y_j \geq 1\), then \( \xi_j^*(w,b) = 0 \)

If \((w \cdot x_j + b) y_j < 1\), then \( \xi_j^*(w,b) = 1 - (w \cdot x_j + b) y_j \)

\[ \xi_j^* = \max \left( 0, 1 - (w \cdot x_j + b) y_j \right) \]
Equivalent hinge loss formulation

\[
\min_{w,b,\xi} \sum_j \xi_j \\
\left( w \cdot x_j + b \right) y_j \geq 1 - \xi_j, \forall j \quad \xi_j \geq 0
\]

Substituting \( \xi_j = \max \left( 0, 1 - (w \cdot x_j + b) y_j \right) \) into the objective, we get:

\[
\min_{w,b} \sum_j \max \left( 0, 1 - (w \cdot x_j + b) y_j \right)
\]

Now an \textit{unconstrained} optimization problem. No longer a \textit{linear} objective, but it is \textit{convex}.
Key idea #2: seek large margin
Key idea #2: seek large margin

- Consider the constraints:
  \[ y_t (w \cdot x_t + b) \geq 1 \quad \forall t \]

- As the norm of the weight vector \(||w||\) and \(b\) get smaller, the optimization problem becomes infeasible:

As \(||w||\) (and \(|b|\)) get smaller
What is $\gamma$ (geometric margin) as a function of $w$?

$\gamma_i =$ Distance to $i$’th data point

$\gamma = \min_i \gamma_i$

We also know that:

$w \cdot x_1 + b = 1$

$w \cdot x_2 + b = 0$

$w \cdot (x_1 - x_2) = 1$

$x_1 - x_2 = \gamma \frac{w}{||w||}$

$1 = w \cdot \left( \gamma \frac{w}{||w||} \right) = \frac{\gamma}{||w||} w \cdot w = \gamma ||w||$

So, $\gamma = \frac{1}{||w||}$ (assuming there is a data point on the $w \cdot x + b = +1$ or $-1$ line)

Final result: can maximize $\gamma$ by minimizing $||w||_2$!!!
(Hard margin) support vector machines

- Example of a convex optimization problem
  - A quadratic program
  - Polynomial-time algorithms to solve!
- Hyperplane defined by support vectors
  - Could use them as a lower-dimension basis to write down line, although we haven’t seen how yet
- More on these later

\[
\begin{align*}
\text{minimize}_{w,b} & \quad w \cdot w \\
& \left(w \cdot x_j + b\right) y_j \geq 1, \quad \forall j
\end{align*}
\]

Non-support Vectors:
- everything else
- moving them will not change \( \mathbf{w} \)

Support Vectors:
- data points on the canonical lines

\[
\mathbf{w} \cdot \mathbf{x} + b = \pm 1
\]
Allowing for slack: “Soft margin SVM”

For each data point:
• If margin $\geq 1$, don’t care
• If margin $< 1$, pay linear penalty

minimize $\mathbf{w}, b$ \[ \mathbf{w}.\mathbf{w} + C \sum_j \xi_j \]
\[ (\mathbf{w}.\mathbf{x}_j + b) y_j \geq 1 - \xi_j, \forall j \quad \xi_j \geq 0 \]
“slack variables”

Slack penalty $C > 0$:
• $C = \infty \rightarrow$ have to separate the data!
• $C = 0 \rightarrow$ ignores the data entirely!
• Select using cross-validation
Equivalent formulation using hinge loss

\[
\min_{w, b} \quad w \cdot w + C \sum_j \xi_j \\
\left( w \cdot x_j + b \right) y_j \geq 1 - \xi_j \quad \forall j \quad \xi_j \geq 0
\]

Substituting \( \xi_j = \max \left( 0, 1 - (w \cdot x_j + b) y_j \right) \) into the objective, we get:

\[
\min ||w||^2 + C \sum_j \max \left( 0, 1 - (w \cdot x_j + b) y_j \right)
\]

Recall, the hinge loss is \( \ell_{\text{hinge}}(y, \hat{y}) = \max \left( 0, 1 - \hat{y}y \right) \)

\[
\min_{w, b} ||w||^2 + C \sum_j \ell_{\text{hinge}}(y_j, w \cdot x_j + b)
\]

This is called regularization; used to prevent overfitting! This part is empirical risk minimization, using the hinge loss