SVM Questions

Lyle Ungar
Hyperplanes

Given the hyperplane defined by the line
- \( y = x_1 - 2x_2 \)
- \( y = (1,-2)^T x = w^T x \)

Is this point correctly predicted?
1) \( y = 1, \quad x = (1,0) \) ?
   - A. Yes
2) \( y = 1, \quad x = (1,1) \) ?
   - B. No
Hyperplanes

Given the hyperplane defined by the line:
- \( y = x_1 - 2x_2 \)
- \( y = (1,-2)^T \cdot x \)

So (1,0) having \( y = 1 \) is correct and (1,1) having \( y = 1 \) is not correct.
The projection of a point $x$ onto a line $w$ is
$$x^T w / \|w\|_2$$

The distance of a point $x$ to a hyperplane defined by the line $w$ is
$$x^T w / \|w\|_2$$

Why?
Hyperplanes

The projection of a point $x$ on a plane defined by a line $w$ is $x^T w / |w|^2$

The distance of $x$ from the hyperplane defined by $(1, -2)$ is what

- for $x = (-1, 2)$
- for $x = (1, 0)$
- for $x = (1, 1)$

A) $1/\sqrt{5}$
B) $-1/\sqrt{5}$
C) $1/5$
D) $\sqrt{5}$
E) None of the above
The projection of a point $x$ on a plane defined by a line $w$ is $x^T w / |w|_2$

The distance of $x$ from the hyperplane defined by $(1, -2)$ is what

- for $x = (-1, 2)$, $\sqrt{5}$
- for $x = (1, 0)$, $1/\sqrt{5}$
- for $x = (1, 1)$, $1/\sqrt{5}$

Project onto the line $x^T (1 -2) / \sqrt{(1^2 + (-2)^2)} = x^T (1 -2) / \sqrt{5}$
Hyperplanes

True or False? When solving for a hyperplane specified by \( \mathbf{w}^\top \mathbf{x} + b = 0 \) one can always set the margin to 1;

\[
\mathbf{w}^\top \mathbf{x}_1 + b = -1 \quad \text{and} \quad \mathbf{w}^\top \mathbf{x}_2 + b = 1
\]
Hyperplanes

True or False? When solving for a hyperplane specified by \( \mathbf{w}^\top \mathbf{x} + b = 0 \) one can always set the margin to 1;

\[
\mathbf{w}^\top \mathbf{x}_1 + b = -1 \quad \text{and} \quad \mathbf{w}^\top \mathbf{x}_2 + b = 1
\]

True (as long as it is separable)
Hyperplanes

◆ **True or False?** When solving for a hyperplane specified by \( w^T x + b = 0 \) one can always set the margin to 1;

\[
w^T x_1 + b = -1 \quad \text{and} \quad w^T x_2 + b = 1
\]

◆ **True or False?** This then implies that the margin, the distance of the support vectors from the separating hyperplane, is

\[
\frac{w^T}{2\|w\|_2} (x_2 - x_1) = \frac{1}{\|w\|_2}
\]
True or False? When solving for a hyperplane specified by \( \mathbf{w}^\top \mathbf{x} + b = 0 \) one can always set the margin to 1;

\[
\mathbf{w}^\top \mathbf{x}_1 + b = -1 \quad \text{and} \quad \mathbf{w}^\top \mathbf{x}_2 + b = 1
\]

True

True or False? This then implies that the margin, the distance of the support vectors from the separating hyperplane, is

\[
\frac{\mathbf{w}^\top}{2\|\mathbf{w}\|_2} (\mathbf{x}_2 - \mathbf{x}_1) = \frac{1}{\|\mathbf{w}\|_2}
\]

True
(non)separable SVMs

- **True or False?** In a real problem, you should check to see if the SVM is separable and then include slack variables if it is not separable.

- **True or False?** Linear SVMs have no hyperparameters that need to be set by cross-validation.
(non)separable SVMs

True or False? In a real problem, you should check to see if the SVM is separable and then include slack variables if it is not separable.

False: you can just run the slack variable problem in either case (but you need to pick C)

True or False? Linear SVMs have no hyperparameters that need to be set by cross-validation

False: you need to pick C

True or False? Adding slack variables is equivalent to requiring that all of the $\alpha_i$ are less than a constant in the dual

$$\mathbf{w}^T \phi(\mathbf{x}) + b = \sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b$$
Non-separable SVMs

True or False? A support vector could be inside the margin.
Non-separable SVMs

- $x_2, x_3$ have $y=1$, while $x_4$ has $y = -1$
- All are co-linear

At what location does $x_3$ become a support vector?

True/False? It just needs to be closer to $x_4$, than $x_2$ is
Consider the following cases
1) a point \( x_1 \) is on the correct side of the margin
2) a point \( x_2 \) is on the margin
3) a point \( x_3 \) is on the wrong side of the margin
4) a point \( x_4 \) is on the wrong side of the decision hyperplane

A) \( \alpha_i = 0 \)
B) \( \alpha_i < C \)
C) \( \alpha_i = C \)

\[ \begin{align*}
\text{Hinge dual:} \\
\max_{\alpha \geq 0} & \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\
\text{s.t.} & \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i \leq C, \quad i = 1, \ldots, n
\end{align*} \]
Non-separable SVMs - dual

Consider the following cases
1) a point $x_1$ is on the correct side of the margin $\alpha_i = 0$ nonbinding
2) a point $x_2$ is on the margin $\alpha_i < C$
3) a point $x_3$ is on the wrong side of the margin $\alpha_i = C$
4) a point $x_4$ is on the wrong side of the decision hyperplane $\alpha_i = C$

If a point is on the wrong side of the margin (cases 3 and 4), $\xi_i > 0$ and hence $\lambda_i = 0$ (last term of the first equation below; $\lambda_i$ is the Lagrange multiplier for the $i$th slack variable $\xi_i$), and hence $\alpha_i = C$

If a point is on the margin (case 2), then $\xi_i = 0$, and $\lambda_i$ can be greater than zero so in general $\alpha_i < C$.

$$L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}w^T w + C \sum_i \xi_i + \sum_i \alpha_i (1 - \xi_i - y_i(w^T x_i + b)) - \sum_i \lambda_i \xi_i$$

$$\frac{\partial L}{\partial \xi_i} = 0 \quad \rightarrow \quad C - \alpha_i - \lambda_i = 0$$
Why do SVMs work well?

- Why are SVMs fast?

- Why are SVMs often more accurate than logistic regression?
Why do SVMs work well?

- **Why are SVMs fast?**
  - Quadratic optimization (convex!)
  - They work in the dual, with relatively few points
  - The kernel trick

- **Why are SVMs often more accurate than logistic regression?**
  - SVMs use kernels – but regression can, too
  - SVMs assume less about the model form
    - Logistic regression uses all the data points, assuming a probabilistic model, while SVMs ignore the points that are clearly correct, and give less weight to ones that are wrong
Hyperplanes

Given the hyperplane defined by the line

- \( y = x_1 - 2x_2 \)
- \( y = (1, -2)^T x = w^T x \)

What is the minimal adjustment to \( w \) to make a new point \( y = 1, x = (1,1) \) be correctly classified?
Hyperplanes

Given the hyperplane defined by the line

\[ y = (1,-2)^T x = w^T x \]

What is the minimal adjustment to \( w \) to make a new point \( y = 1, x = (1,1) \) be correctly classified?

Make \( w = (1,-1) \)