Lecture 35, Apr 10, 2023

Support Vector Machines

- In normal regression we had $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$; for classification problems, they generally have input $x \in \mathbb{R}^n$ with output $y \in \{-1, 1\}$, i.e. the output is a binary yes or no
- We can express a hyperplane as $\boldsymbol{w}^T \boldsymbol{x} - b = 0$, where $\boldsymbol{w}$ is the normal vector defining the orientation and b is an offset from the origin
- The hyperplane divides all of the input space into 2 regions, $\boldsymbol{w}^T \boldsymbol{x} > b$ and $\boldsymbol{w}^T \boldsymbol{x} < b$; each region corresponds to a different value of y
- Given some data, we're looking for a hyperplane that separates the 2 types of data
  - We also want a hyperplane that's the most "in the middle" and divides the empty space between 2 types evenly
- We want to find $\boldsymbol{w}^T \boldsymbol{x} - b = 0$ that maximizes d, the distance on each side of the hyperplane, while separating the data
- Unlike in linear regression, this problem is not analytically solvable
  - In linear regression, a change in any data point is going to affect the total error and therefore change the solution; however in this problem moving a data point may not affect the solution at all
  - Depending on the orientation of the hyperplane $\boldsymbol{w}$, different data points will become relevant
- What is the expression for d?
  - Consider 2 parallel hyperplanes, $\boldsymbol{w}^T \boldsymbol{x} = 1$ and $\boldsymbol{w}^T \boldsymbol{x} = 0$ and some point on the first hyperplane so $\boldsymbol{w}^T \boldsymbol{x}^* = 1$, then $d = \|\boldsymbol{x}^*\|$
  - We know $\boldsymbol{x}^* = \alpha \boldsymbol{w}$ so $\|\boldsymbol{x}^*\| = \dfrac{1}{\|\boldsymbol{w}\|}$, $\alpha = \dfrac{1}{\|\boldsymbol{w}\|^2} \implies \boldsymbol{x}^* = \dfrac{\boldsymbol{w}}{\|\boldsymbol{w}\|^2}$
  - $d = \dfrac{1}{\|\boldsymbol{w}\|}$
- To maximize d we want to minimize $\|\boldsymbol{w}\|$ or $\|\boldsymbol{w}\|^2$, subject to the constraint that if $y_i = 1$, then $\boldsymbol{w}^T \boldsymbol{x}_i - b > 0$, or if $y_i = -1$, then $\boldsymbol{w}^T \boldsymbol{x}_i - b < 0$
- The support vector machine is $\min_{\boldsymbol{w}, b} \|\boldsymbol{w}\|^2$ such that $y_i(\boldsymbol{w}^T \boldsymbol{x}_i - b) \geq 0$ for all training data
  - This is a quadratic program
  - If this is solvable, i.e. the data is separable, then we have an optimal classifier