CS 189: Introduction to Machine Learning - Discussion 5

1. Fun with Newton's method for root-finding

   (a) Write down the iterative update equation of Newton's method for finding a root $x : f(x) = 0$ for a real-valued function f.

   (b) Prove that if $f(x)$ is a quadratic function ($f(x) = ax^2 + bx + c$), then it only takes one iteration of Newton's Method to find the minimum/maximum.

2. Linearly Separable Data with Logistic Regression

   Show (or explain) that for a linearly separable data set, the maximum likelihood solution for the logistic regression model is obtained by finding a vector β whose decision boundary $\beta^T x = 0$ separates the classes, and taking the magnitude of β to be infinity. **Note**: Remember that as mentioned in lecture, doing maximum-likelihood on logistic regression is same as minimizing cross-entropy loss (see lecture-6, slides-21,22). In lecture, we explored the cross-entropy loss-minimization perspective to logistic regression. This question will make you explore the likelihood perspective.

3. Linear Regression with Laplace prior

   We saw in discussion 4 that there is a probabilistic interpretation of linear regression: $P(y|\mathbf{x}, \sigma^2) \sim \mathcal{N}(\mathbf{w^T x}, \sigma^2)$. We extend this by assuming some prior distribution on parameters $\mathbf{w}$. Let us assume the prior is a Laplace distribution, so we have:

$$w_j \sim Laplace(0, t), \text{ i.e. } P(w_j) = \frac{1}{2t} e^{-|w_j|/t} \text{ and } P(\mathbf{w}) = \prod_{j=1}^{D} P(w_j) = (\frac{1}{2t})^D \cdot e^{-\frac{\sum |w_j|}{t}}$$

   Show it is equivalent to minimizing the following risk function, and find the value of the constant λ:

$$R(\mathbf{w}) = \sum_{i=1}^{n} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \lambda \|\mathbf{w}\|_1, \text{ where } \|\mathbf{w}\|_1 = \sum_{j=1}^{D} |w_j|$$

4. Review: Linear SVM in Higher Dimensional space (video)
   Consider a data set, $X \in \mathbb{R}^{nxd}$.
   Let $X_i \in \mathbb{R}^d$ be one data point, i.e. one row of X. We can create a quadratic feature vector $X_i^{'}$ from X_i by mapping the features:
   $x_1, x_2, ..., x_d$ to
   $x_1^2, x_2^2, ...x_d^2, \sqrt{2}x_1x_2, ..., \sqrt{2}x_1x_d, \sqrt{2}x_2x_1, .., \sqrt{2}x_2x_d, ...\sqrt{2}x_{d_1}x_d$.

   For simplicity, lets consider the simple case where our data is initially two dimensional: A quadratic mapping takes x_1, x_2 to $x_1^2, x_2^2, \sqrt{2}x_1x_2$.
   We can view these terms as a new feature vector, and fit a linear decision boundary in this higher, $3D$ space. The boundary will be linear in the features.

   This can also be viewed as fitting a polynomial boundary in a (d+1) dimensional space.

   The following video demonstrates this concept: `https://www.youtube.com/watch?v=3liCbRZPrZA`