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.

   Solution: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

   (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.

   Solution: The Newton's method update for finding a minimum/maximum is

   $$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} = x_n - \frac{2ax_n + b}{2a} = \frac{-b}{2a}$$

   And this is the point for 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.

   Solution:

   Because the data is linearly separable, it is possible to find a hyperplane with unit normal vector β such that each halfspace induced by this hyperplane contain all samples of one class.

   Consider all points on the half space defined by $\beta^T x \geq 0$. Without loss of generality, let's say that all these points come from class 1, while the points such that $\beta^T x < 0$ come from class 0. For some point x_{c_1} in class 1,

   $$P(y = 1|x_{c_1}) = \mu_1 = \frac{1}{1 + exp(-\beta^T x_{c_1})} > 0.5$$

because $\beta^T x_{c_1} \geq 0$. Likewise, for a point x_{c_0} in class 0,

$$P(y = 0|x_{c_0}) = 1 - P(y = 1|x_{c_0}) = 1 - \mu_1 > 0.5$$

since $\beta^T x_{c_0} < 0$. Now, when we inspect the likelihood of the data, given by

$$L(\beta|D) = \prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i} = \prod_{i \in C_1} \mu_i \prod_{j \in C_0} (1 - \mu_j)$$

we see that if we take some arbitrary $k > 1$ and scale the unit vector β by k, our likelihood will increase, since all of the individual probabilities in the likelihood will increase. In fact, we can set $k = \infty$, which will maximize our likelihood. This will render the sigmoid function to be infinitely steep at $\beta^T x_i = 0$ (making it a step function). $P(y = y_i|x_i) = 1$ for all x_i, and the likelihood will be 1. Obviously this is severely overfitting the data, and regularization for this problem would help us avoid that issue.

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:

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

Solution: Note that $\mathbf{X_i} = \mathbf{x}^{(i)}, Y_i = y^{(i)}$. We have to solve the MAP for parameter $\mathbf{w}$ and the posterior of $\mathbf{w}$ is,

$$P(w|\mathbf{X_i}, Y_i) \propto (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot P(\mathbf{w}) = (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot \prod_{j=1}^{D} P(w_j)$$

Taking log and we want to maximize

$$\begin{aligned} l(\mathbf{w}) &= \sum_{i=1}^{n} log\mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2) + \sum_{j=1}^{D} logP(w_j) \\ &= \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2})) + \sum_{j=1}^{D} log(\frac{1}{2t} exp(\frac{-|w_j|}{t})) \\ &= -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2} + \frac{-\sum_{j=1}^{D} |w_j|}{t} + nlog(\frac{1}{\sqrt{2\pi}\sigma}) + Dlog(\frac{1}{2t}) \end{aligned}$$

So it is equivalent to minimizing the following function:

$$R(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \frac{2\sigma^2}{t} \sum_{j=1}^{D} |w_j| = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \|\mathbf{w}\|_1$$

where $\lambda = \frac{2\sigma^2}{t}$.

This form of linear regression is called *ridge* regression.

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`