CSE – 426 Homework 6

Griffin Kent

1) Prove the following solutions to the $\ell 2$ and $\ell 1$ regularization optimization problems with a single parameter $w \in \mathbb{R}$. x and $\lambda \geq 0$ are two constant scalars.

- $\ell 2$ regularized optimization

$$\min_w \frac{1}{2} w^2 - xw + \lambda \|w\|^2.$$

The optimal solution is $w^* = \frac{x}{1+2\lambda}$, which is not zero when $x \neq 0$.

- $\ell 1$ regularized optimization

$$\min_w \frac{1}{2} w^2 - xw + \lambda |w|$$

The optimal solution is $w^* = \text{sign}(x)\max\{0, |x| - \lambda\}$, where the $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$, and $\text{sign}(0) = 0$. This solution will be 0 if $|x| < \lambda$ so that x^* will be 0 if $|x|$ is not significant compared to λ.

[*Hints: Since the $\ell 2$ regularized problem is an unconstrained, convex, and single variable optimization problem, you can simply set the derivative of the objective function with respect to w to 0 and solve for w. The $\ell 1$ regularized problem is more difficult. You need to be careful about the derivative of the function $|w|$, which is not differentiable at $w = 0$. You should take the following route: for the four cases when $x - \lambda < 0$, $x - \lambda > 0$, $x + \lambda < 0$, and $x + \lambda > 0$, respectively, solve for the optimal w, and then prove that all the cases lead to the same formula $w^* = \text{sign}(x)\max\{0, |x| - \lambda\}$.*]

(Proofs):

- **$\ell 2$ regularized optimization**

Since the $\ell 2$ regularized problem is unconstrained, convex, and a single-variable optimization problem, we will take the derivative of the objective function with respect to w, set it equal to 0 and solve for w^*. Taking the derivative of the objective function

$$F(w) = \frac{1}{2} w^2 - xw + \lambda \|w\|^2$$

We have

$$\nabla_w \mathrm{F} = \frac{\partial F}{\partial w} = w - x + 2\lambda w$$

Setting to 0 and solving:

$$w - x + 2\lambda w = 0$$
$$\rightarrow w^* = \frac{x}{1 + 2\lambda}.$$

■

- **$\ell 1$ regularized optimization**

To start, we can get rid of the absolute value by looking at the two cases when w is positive and negative.

Case $+w$:

We have $\frac{1}{2}w^2 - xw + \lambda w = \frac{1}{2}w^2 + w(\lambda - x)$. Using the quadratic formula, we have

$$w = (x - \lambda) \pm (\lambda - x) = 0, 2x - 2\lambda.$$

Thus, we can see that $w^* = \begin{cases} 0, & \text{if } x - \lambda < 0 \\ x - \lambda, & \text{if } x - \lambda > 0 \end{cases}$, which is equivalent to $w^* = \text{sign}(x)\max\{0, |x| - \lambda\}$.

Case $-w$:

We have $\frac{1}{2}w^2 + xw - \lambda w = \frac{1}{2}w^2 + w(x - \lambda)$. Using the quadratic formula, we have

$$w = (\lambda - x) \pm (x - \lambda) = 0, 2\lambda - 2x.$$

Thus, we can see that $w^* = \begin{cases} 0, & \text{if } \lambda - x < 0 \\ \lambda - x, & \text{if } \lambda - x > 0 \end{cases} \rightarrow \begin{cases} 0, & \text{if } x - \lambda > 0 \\ \lambda - x, & \text{if } x - \lambda < 0 \end{cases}$, which is equivalent to $w^* = \text{sign}(x)\max\{0, |x| - \lambda\}$.

Therefore, since every case has resulted in the same result, we have shown that the optimal solution is $w^* = \text{sign}(x)\max\{0, |x| - \lambda\}$.

■

2) For the joint probability distribution

$$\mathbb{P}(x = 0, y = 0) = 0.1$$
$$\mathbb{P}(x = 0, y = 1) = 0.3$$
$$\mathbb{P}(x = 1, y = 0) = 0.4$$
$$\mathbb{P}(x = 1, y = 1) = 0.2$$

Compute the Mutual Information between the two random variables x and y.

(Solution): The Mutual Information between two random variables x and y is defined as

$$MI(X, Y) = \sum_x \sum_y \mathbb{P}(x, y) \log \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)}$$

Where the summation is over all the possible values of x and y. From the joint probabilities given above, we can find the marginal probabilities to be the following:

$$\mathbb{P}(x = 0) = \mathbb{P}(x = 0, y = 0) + \mathbb{P}(x = 0, y = 1) = 0.1 + 0.3 = 0.4.$$
$$\mathbb{P}(x = 1) = \mathbb{P}(x = 1, y = 0) + \mathbb{P}(x = 1, y = 1) = 0.4 + 0.2 = 0.6.$$
$$\mathbb{P}(y = 0) = \mathbb{P}(x = 0, y = 0) + \mathbb{P}(x = 1, y = 0) = 0.1 + 0.4 = 0.5.$$
$$\mathbb{P}(y = 1) = \mathbb{P}(x = 0, y = 1) + \mathbb{P}(x = 1, y = 1) = 0.3 + 0.2 = 0.5.$$

Now, we can calculate the Mutual Information to be:

$$MI(X, Y)$$
$$= \mathbb{P}(x = 0, y = 0) \log \frac{\mathbb{P}(x = 0, y = 0)}{\mathbb{P}(x = 0)\mathbb{P}(y = 0)}$$
$$+ \mathbb{P}(x = 0, y = 1) \log \frac{\mathbb{P}(x = 0, y = 1)}{\mathbb{P}(x = 0)\mathbb{P}(y = 1)}$$
$$+ \mathbb{P}(x = 1, y = 0) \log \frac{\mathbb{P}(x = 1, y = 0)}{\mathbb{P}(x = 1)\mathbb{P}(y = 0)}$$
$$+ \mathbb{P}(x = 1, y = 1) \log \frac{\mathbb{P}(x = 1, y = 1)}{\mathbb{P}(x = 1)\mathbb{P}(y = 1)}$$

$$= (0.1)\log\frac{0.1}{(0.4)(0.5)} + (0.3)\log\frac{0.3}{(0.4)(0.5)} + (0.4)\log\frac{0.4}{(0.6)(0.5)} + (0.2)\log\frac{0.2}{(0.6)(0.5)}$$
$$\approx 0.03748.$$
■

3) Failure of feature selection. Let the feature x_1 be a random variable distributed uniformly in the interval $(-1,1)$ and the target variable be calculated as $y = x_1^2$ deterministically. Let the feature x_2 be determined stochastically by $x_2 = y + z = x_1^2 + z$, where z is a random variable distributed uniformly in the interval $(-0.01,0.01)$. If x_1 (not x_1^2) is selected to predict y, (that is, $\hat{y} = x_1$) what is the largest deviation between the predicted $\hat{y}$ and the true y? If x_2 is selected instead to predict y, what is the largest deviation then? The results demonstrate that selecting the feature that deterministically defines the target may not be optimal if the predictor does not use the selected feature correctly.

(Solution): Given that x_1 is a uniformly distributed random variable over the interval $(-1,1)$, we can see that the greatest deviation between the true value of $y = x_1^2$ versus $\hat{y} = x_1$ will be when $x_1 = -1$; this will yield a maximum deviation of

$$|\hat{y} - y| = |x_1 - x_1^2| = |-1 - 1| = 2.$$

Now, if x_2 were selected to predict y such that $\hat{y} = x_2 = x_1^2 + z$, then we can see that the greatest deviation could occur when $x_1 = \pm 1$ and $z = -0.01$; this will yield a maximum deviation of

$$|\hat{y} - y| = |x_1^2 + z - x_1^2| = |1 + (-0.01) - 1| = 0.01.$$

■

4) K-means on the real line. Given the unlabeled training data $x^{(1)} = 0$, $x^{(2)} = 1$, $x^{(3)} = -2$, find the two centers μ_1 and μ_2 on the real line and the corresponding assignment r_{ik} of the training data to the two clusters, so that the distortion measure $J = \sum_{i=1}^{3}\sum_{k=1}^{2} r_{ik}\left(x^{(i)} - \mu_k\right)^2$ is minimized. What is the smallest J value?

[*Hints: There are only a finite number of possible assignments with different μ_k, and you can exhaust them and select the optimal assignment.*]

(Solution): To begin, let's assign $x^{(1)}$ and $x^{(2)}$ to center μ_1 and $x^{(3)}$ to center μ_2. Then we have the coordinate of the centers:

$$\mu_1 = \frac{\sum_{i=1}^{m} r_{i1}x^{(i)}}{\sum_{i=1}^{m} r_{i1}} = \frac{(1)(0) + (1)(1) + (0)(-2)}{2} = \frac{1}{2}.$$
$$\mu_2 = \frac{\sum_{i=1}^{m} r_{i2}x^{(i)}}{\sum_{i=1}^{m} r_{i2}} = \frac{(0)(0) + (0)(1) + (1)(-2)}{1} = -2.$$

Calculating the distortion measure for these centers:

$$J = \left(0 - \frac{1}{2}\right)^2 + \left(1 - \frac{1}{2}\right)^2 + (-2 - (-2))^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{5}.$$

Performing the same calculations for $x^{(1)}$ in cluster μ_1 and $x^{(2)}$ and $x^{(3)}$ in cluster μ_2:

$$\mu_1 = x^{(1)} = 0.$$
$$\mu_2 = \frac{x^{(2)} + x^{(3)}}{2} = \frac{1 + (-2)}{2} = -\frac{1}{2}.$$
$$J = (0-0)^2 + \left(1 - \left(-\frac{1}{2}\right)\right)^2 + \left(-2 - \left(-\frac{1}{2}\right)\right)^2 = \frac{9}{4} + \frac{9}{4} = \frac{9}{2}.$$

Performing the same calculations for $x^{(1)}$ and $x^{(3)}$ in cluster μ_1 and $x^{(2)}$ in cluster μ_2:

$$\mu_1 = \frac{x^{(1)} + x^{(3)}}{2} = \frac{0 + (-2)}{2} = -1.$$
$$\mu_2 = x^{(2)} = 1.$$
$$J = \big(0 - (-1)\big)^2 + (1-1)^2 + \big(-2 - (-1)\big)^2 = 1 + 1 = 2.$$

Since we have exhausted the possible clustering solutions of the three training examples, we can see that the clustering that minimizes the distortion measure will be centers $\mu_1 = \frac{1}{2}$ and $\mu_2 = -2$ with $x^{(1)}$ and $x^{(2)}$ in cluster 1 and $x^{(3)}$ in cluster 2.

■

5) (Graduate Only) Prove that the MAP (maximum a posterior) of Bayesian linear regression, presented in section 6.2 of the learning theory notes, leads to the $\ell 2$ regularized linear regression problem

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{m}\big(\boldsymbol{w}^T\boldsymbol{x}^{(i)} - y^{(i)}\big)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2.$$

[*Hints: Take the log of the posterior in Eq. (41) of the learning theory lecture notes and simplify to obtain the above regression problem.*]

(Proof): The MAP of w is defined as

$$\boldsymbol{w}^*_{MAP} = \arg\max_{\boldsymbol{w}\in\mathbb{R}^n} \mathbb{P}\left(\boldsymbol{w}\middle|\left\{\boldsymbol{x}^{(i)}, y^{(i)}\right\}_{i=1}^{m}; \tau, \sigma\right)$$
$$= \mathbb{P}(\boldsymbol{w};\tau)\prod_{i=1}^{m}\mathbb{P}\big(y^{(i)}\big|\boldsymbol{x}^{(i)}, \boldsymbol{w}; \sigma\big).$$

Since we know that $\boldsymbol{w}\sim N(0, \tau^2 I)$, we know that

$$\mathbb{P}(\boldsymbol{w};\tau) = \frac{1}{(2\pi)^{n/2}|\tau^2\boldsymbol{I}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{w})^T(\tau^2\boldsymbol{I})^{-1}(\boldsymbol{w})}.$$

Likewise, since we know that $y^{(i)}\sim N\big(\boldsymbol{w}^T\boldsymbol{x}^{(i)}, \sigma^2\big)$, we know that

$$\mathbb{P}\big(y^{(i)}\big|\boldsymbol{x}^{(i)}, \boldsymbol{w}; \sigma\big) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2/(2\sigma^2)}$$

Then, taking the log of the MAP of w, we have

$$\log(\boldsymbol{w}^*_{MAP}) = \log\left(\mathbb{P}(\boldsymbol{w};\tau)\prod_{i=1}^{m}\mathbb{P}\big(y^{(i)}\big|\boldsymbol{x}^{(i)}, \boldsymbol{w}; \sigma\big)\right)$$
$$= \log\big(\mathbb{P}(\boldsymbol{w};\tau)\big) + \sum_{i=1}^{m}\log\left(\mathbb{P}\big(y^{(i)}\big|\boldsymbol{x}^{(i)}, \boldsymbol{w}; \sigma\big)\right)$$

$$= \log\left(\frac{1}{(2\pi)^{n/2}|\tau^2\boldsymbol{I}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{w})^T(\tau^2\boldsymbol{I})^{-1}(\boldsymbol{w})}\right) + \sum_{i=1}^{m} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2/(2\sigma^2)}\right)$$

$$= \left[-\frac{1}{2}(\boldsymbol{w})^T(\tau^2\boldsymbol{I})^{-1}(\boldsymbol{w})\right]\log\left(\frac{1}{(2\pi)^{n/2}|\tau^2\boldsymbol{I}|^{1/2}}\right) + \sum_{i=1}^{m}\left\{\left[-\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2/(2\sigma^2)\right]\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\right\}$$

$$= \left[-\frac{1}{2}(\boldsymbol{w})^T\left(\frac{\boldsymbol{I}}{\tau^2}\right)(\boldsymbol{w})\right]\log\left(\frac{1}{(2\pi)^{n/2}\tau^n}\right) + \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\left(-\frac{1}{2\sigma^2}\right)\sum_{i=1}^{m}\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2$$

$$= \left[-\frac{1}{2\tau^2}(\boldsymbol{w})^T(\boldsymbol{I})(\boldsymbol{w})\right]\log\left(\frac{1}{(2\pi)^{n/2}\tau^n}\right) + \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\left(-\frac{1}{2\sigma^2}\right)\sum_{i=1}^{m}\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2$$

$$= -\frac{\log\left(\frac{1}{(2\pi)^{\frac{n}{2}}\tau^n}\right)}{2\tau^2}\|\boldsymbol{w}\|^2 + \left(-\frac{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}{\sigma^2}\right)\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2$$

$$= \frac{\beta}{2}\sum_{i=1}^{m}\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \frac{\alpha}{2}\|\boldsymbol{w}\|^2.$$

Where $\alpha \triangleq -\frac{\log\left(\frac{1}{(2\pi)^{n/2}\tau^n}\right)}{\tau^2}$ and $\beta \triangleq -\frac{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)}{\sigma^2}$.

Finally, letting $\lambda \triangleq \frac{\alpha}{\beta}$, and multiplying by $\frac{1}{\beta}$, we have the result

$$\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

Therefore, we can see that

$$\boldsymbol{w}^*_{MAP} = \arg\max_{w\in\mathbb{R}^n} \mathbb{P}\left(\boldsymbol{w}\middle|\left\{\boldsymbol{x}^{(i)}, y^{(i)}\right\}_{i=1}^{m}; \tau, \sigma\right)$$

Is equivalent to

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2.$$

■