Homework 6

Alex Pine

2015/04/09

1 Robust Ridge Regression

1.1 Write the objective function for ℓ_2-regularized emprical risk minimization with an ℓ_1 loss, over a linear hypothesis space.

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} |x_i^T - y_i| + \lambda ||w||^2$$

1.2 Show that the objective function is convex

$x_i^T - y_i$ is an affine mapping, and the L_1 norm is convex. Therefore $|x_i^T - y_i|$ is convex, because it is a convex function composed with an affine mapping.

$\frac{1}{n} \sum_{i=1}^{n} |x_i^T - y_i|$ is convex because it is the sum of weighted convex functions.

$||w||$ is convex because it is a norm function. $||w||^2$ is convex because $||w||$ is positive, and a polynomial of a convex function is convex if the domain of the function is positive.

Lastly, $\frac{1}{n} \sum_{i=1}^{n} |x_i^T - y_i| + \lambda ||w||^2$ is convex because it is the sum of two convex functions.

1.3 Write the objective function as a quadratic program (i.e. an optimization problem with a quadratic objective and linear constraints).

$J(w) = \frac{1}{n} \sum_{i=1}^{n} |x_i^T - y_i|$, subject to $||w||^2 \leq r, r \geq 0$

2 RBF Kernel

2.1 Show that the distance between the feature representations of any two points x_i and x_j in the space H is at most $\sqrt{2}$.

Proof my contradiction.

Assert $d(\phi(x_i), \phi(x_j)) > \sqrt{2}$, then

$$\begin{aligned}
\|\phi(x_i), \phi(x_j)\| &> \sqrt{2} \\
\langle \phi(x_i) - \phi(x_j), \phi(x_i) - \phi(x_j) \rangle &> 2 \\
\langle \phi(x_i), \phi(x_i) \rangle + \langle \phi(x_j), \phi(x_j) \rangle - 2\langle \phi(x_i), \phi(x_j) \rangle &> 2 \\
2 - 2k(x_i, x_j) &> 2 \\
k(x_i, x_j) &< 0 \\
\exp(-\frac{1}{2}\|x_i - x_j\|^2) < 0 &\implies \text{impossible!} \\
\therefore d(\phi(x_i), \phi(x_j)) &\leq \sqrt{2}
\end{aligned}$$

3 Regularized Logistic Regression

3.1 J(w) has multiple locally optimal solutions? T/F?

False, J is convex, so it has one optimal solution. It's convex because its loss function and regularization term are convex. Its loss function is convex because the negative logarithm of a positive function is convex, and the sigmoid function is a positive function. Additionally, the regularization term is convex since it's a polynomial of a norm.

3.2 Let $\hat{w} = \text{argmin}_w J(w)$ be a global optimum. Then $\hat{w}$ is sparse (i.e. has many zero entries). T/F?

False, L_2 regularization does not lead to sparse solutions.

3.3 . [Optional] If the training data are linearly separable, then some weights w_j might become infinite if λ = 0. T/F?

True, if there is no regularization term, there is nothing to prevent the weights from growing to increase the margin at each iteration of optimization..

3.4 $L(w, D_{\text{train}})$ always increases as we increase λ. [NOTE: L is the log-likelihood, and the negative empirical risk.] T/F?

False, $L(w, D_{\text{train}})$ will DECREASE as lambda increases. As lambda increases, J can only be maintained by decreasing the L_2 norm of w. Since an unconstrained w maximizes $L(w, D_{\text{train}})$, changing w must decrease it.

3.5 $L(w, D_{\text{test}})$ always increases as we increase λ. T/F?

False. This is typically true, but it depends on the test set. For example, if the test set is identical to the training set with a tiny term added to each datapoint, a zero lambda will lead to a w that causes $L(w, D_{\text{test}})$ to be larger than a large lambda would.

4 Neural Networks with Linear Activation Function

4.1 Redesign the neural network to compute the same function without using any hidden units. Draw the equivalent network and give expressions detailing for the new weights in terms of the old weights and the constant c.

$w_a = c(w_1 w_5 + w_2 w_6)$

$w_b = c(w_3 w_5 + w_4 w_6)$

$y = w_a x_1 + w_b x_2$

4.2 Can the space of functions that is represented by the above neural network also be represented by linear regression?

Yes. After simplifying it, the decisino function is a linearly combination of the elements of x, which is exactly what linear regression does.

In [ ]: