Stanford CS 229, Public Course, Problem Set 2

Dylan Price

December 24, 2016

1

a)

Find a closed-form expreesion for the value of θ which minimizes the ridge regression cost function:
$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 + \frac{\lambda}{2}\|\theta\|^2$

First, put $J(\theta)$ into matrix notation

$$\begin{aligned}
J(\theta) &= \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 + \frac{\lambda}{2}\|\theta\|^2 \\
&= \frac{1}{2}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})(\theta^T x^{(i)} - y^{(i)}) + \frac{\lambda}{2}\|\theta\|^2 \\
&= \frac{1}{2}(X\theta - \vec{y})(X\theta - \vec{y}) + \frac{\lambda}{2}\theta^T\theta \\
&\text{(where } X \text{ is the design matrix and } \vec{y} \text{ is the vector of target values)} \\
&= \frac{1}{2}((X\theta)^T X\theta - (X\theta)^T\vec{y} - \vec{y}^T X\theta + \vec{y}^T\vec{y})) + \frac{\lambda}{2}\theta^T\theta \\
&= (\frac{1}{2}\theta^T X^T X\theta - \frac{1}{2}\theta^T X^T\vec{y} - \frac{1}{2}\vec{y}^T X\theta + \frac{1}{2}\vec{y}^T\vec{y}) + \frac{\lambda}{2}\theta^T\theta
\end{aligned}$$

Now find the gradient

$$\begin{aligned}
\nabla_\theta J(\theta) &= X^T X\theta - \frac{1}{2}X^T\vec{y} - \frac{1}{2}X^T\vec{y} + \lambda\theta \\
&= X^T X\theta - X^T\vec{y} + \lambda\theta
\end{aligned}$$

Set the gradient equal to 0 and solve for θ

$$\begin{aligned}
X^T X\theta - X^T\vec{y} + \lambda\theta &= 0 \\
X^T X\theta + \lambda\theta &= X^T\vec{y} \\
(X^T X + \lambda I)\theta &= X^T\vec{y} \\
\theta &= (X^T X + \lambda I)^{-1} X^T\vec{y}
\end{aligned}$$

b)

Suppose that we want to use kernels to implicitly represent our feature vectors in a high-dimensional (possibly infinite dimensional) space.
Making a prediction on a new input x_{new} would now be done by computing $\theta^T \phi(x_{new})$.
Show how we can use the "kernel trick" to obtain a closed form for the prediction on the new input without ever explicitly computing $\phi(x_{new})$.

$$\begin{aligned} h_\theta(x_{new}) &= \theta^T x_{new} \\ &= ((X^T X + \lambda I)^{-1} X^T \vec{y})^T x_{new} \\ &= ((X X^T + \lambda I)^{-1} \vec{y})^T X x_{new} \\ &= \vec{y}^T (X X^T + \lambda I)^{-T} X x_{new} \end{aligned}$$

Now we replace $x^{(i)}$ with $\phi(x^{(i)})$ for $i = 1...m$,
define kernel function $K(x, z) = \phi(x)\phi(z)$
and kernel matrix $K \in \mathbb{R}^{mxm}$ such that $K_{ij} = K(x^{(i)}, x^{(j)})$

$$h_\theta(x_{new}) = \vec{y}^T (K + \lambda I)^{-T} \sum_{i=1}^{m} K(x^{(i)}, x_{new})$$

2

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \xi_i^2 \\ \text{subject to} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, i = 1, ..., m \end{aligned}$$

a)

Notice that we have dropped the $\xi_i \geq 0$ constraint in the ℓ_2 problem. Show that these non-negativity constraints can be removed. That is, show that the optimal value of the objective will be the same whether or not these constraints are present.

Let ξ_i be a negative error in $\xi_0, ..., \xi_m$. Since ξ_i^2 is positive, ξ_i will contribute the same amount to the objective as $-\xi_i$. Therefore the minimization of the objective does not depend on the sign of ξ_i and the $\xi_i \geq 0$ constraint is irrelevant.

b)

What is the Lagrangian of the ℓ_2 soft margin SVM optimization problem?
Let

$$\begin{aligned} f(w) &= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \xi_i^2 \\ g_i(w) &= -y^{(i)}(w^T x^{(i)} + b) + 1 - \xi_i \end{aligned}$$

Our optimization problem is

$$\begin{aligned} \min \quad & f(w) \\ \text{subject to} \quad & g_i(w) \geq 0, i = 1, ..., m \end{aligned}$$

The Lagrangian is

$$\begin{aligned} \mathcal{L}(w, b, \alpha, \xi) &= f(w) + \sum_{i=1}^{m} \alpha_i g_i(w) \\ &= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \xi_i^2 + \sum_{i=1}^{m} \alpha_i(-y^{(i)}(w^T x^{(i)} + b) + 1 - \xi_i) \\ &= \frac{1}{2}w^T w + \frac{C}{2}\xi^T \xi + \sum_{i=1}^{m} \alpha_i(-y^{(i)} w^T x^{(i)} - y^{(i)} b + 1 - \xi_i) \\ &= \frac{1}{2}w^T w + \frac{C}{2}\xi^T \xi - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \xi_i \end{aligned}$$

c)

Minimize the Lagrangian with respect to w, b, and ξ by taking the following gradients: $\nabla_w \mathcal{L}$, $\frac{\partial \mathcal{L}}{\partial b}$, and $\nabla_\xi \mathcal{L}$, and then setting them equal to 0. Here $\xi = |\xi_1, \xi_2, ..., \xi_m|^T$.

First find $\nabla_w \mathcal{L}$

$$\begin{aligned} \nabla_w \mathcal{L} &= w + \sum_{i=1}^{m} -\alpha_i y^{(i)} x^{(i)} \\ &= w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \end{aligned}$$

Set equal to 0

$$\begin{aligned} 0 &= w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \\ w &= \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \end{aligned}$$

Next find $\frac{\partial \mathcal{L}}{\partial b}$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{m} -\alpha_i y^{(i)}$$

Set equal to 0

$$0 = \sum_{i=1}^{m} -\alpha_i y^{(i)}$$
$$0 = \sum_{i=1}^{m} \alpha_i y^{(i)}$$

Find $\nabla_\xi \mathcal{L}$

$$\nabla_\xi \mathcal{L} = C\xi - \alpha$$

Set equal to 0

$$0 = C\xi - \alpha$$
$$\xi = \frac{1}{C}\alpha$$

Plug these results back into the Lagrangian

$$\begin{aligned}
\mathcal{L}(w, b, \alpha, \xi) &= \frac{1}{2} w^T w + \frac{C}{2}\xi^T \xi - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \xi_i \\
\min_{w,b,\xi} \quad \mathcal{L} &= \frac{1}{2}(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)})^T (\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}) + \frac{C}{2}(\frac{1}{C}\alpha)^T (\frac{1}{C}\alpha) \\
&\quad - \sum_{i=1}^{m} \alpha_i y^{(i)} (\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)})^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i (\frac{1}{C}\alpha_i) \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \frac{1}{2C}\alpha^T \alpha \\
&\quad - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \frac{1}{C}\alpha_i^2 \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \frac{1}{2C} \sum_{i=1}^{m} \alpha_i^2 \\
&\quad - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i - \frac{1}{C} \sum_{i=1}^{m} \alpha_i^2 \\
&= -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - b \sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i - \frac{1}{2C} \sum_{i=1}^{m} \alpha_i^2
\end{aligned}$$

From above we know that $0 = \sum_{i=1}^{m} \alpha_i y^{(i)}$ when the Lagrangian is minimized, so we end with

$$\min_{w,b,\xi} \quad \mathcal{L} = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \sum_{i=1}^{m} \alpha_i - \frac{1}{2C} \sum_{i=1}^{m} \alpha_i^2$$

d)

What is the dual of the ℓ_2 soft margin SVM optimization problem?

$$\begin{aligned} \max_{\alpha} \quad & \theta_D \\ \text{subject to} \quad & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0, i = 1, ..., m \end{aligned}$$

$$\text{where} \quad \theta_D = \min_{w,b,\xi} \quad \mathcal{L}, \text{ which we found above}$$

3

a)

$f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} K(x^{(i)}, x) + b \quad ; \quad K(x, z) = exp(\frac{-||x - z||^2}{\tau^2})$

Find values for the set of parameters $\{\alpha_1, ..., \alpha_m, b\}$ and Gaussian kernel width τ such that $x^{(i)}$ is correctly classified for all $i = 1, ..., m$.

Let $\alpha_i = 1, b = 0$ for $i = 1, ..., m$. Then

$f(x) = \sum_{i=1}^{m} y^{(i)} e^{\frac{-||x^{(i)} - x||^2}{\tau^2}}$

Find τ such that $|f(x^{(i)}) - y^{(i)}| < 1$ for all i.

$$\begin{aligned} |f(x^{(i)}) - y^{(i)}| &= |\sum_{j=1}^{m} y^{(j)} e^{\frac{-||x^{(j)} - x^{(i)}||^2}{\tau^2}}| - y^{(i)}| \\ &= |y^{(i)} e^{\frac{-||x^{(i)} - x^{(i)}||^2}{\tau^2}} + \sum_{j \neq i} y^{(j)} e^{\frac{-||x^{(j)} - x^{(i)}||^2}{\tau^2}} - y^{(i)}| \\ &= |\sum_{j \neq i} y^{(j)} e^{\frac{-||x^{(j)} - x^{(i)}||^2}{\tau^2}}| \\ &\leq \sum_{j \neq i} |y^{(i)} e^{\frac{-||x^{(j)} - x^{(i)}||^2}{\tau^2}}| && \text{By the triangle inequality} \\ &= \sum_{j \neq i} |e^{\frac{-||x^{(j)} - x^{(i)}||^2}{\tau^2}}| && y^{(i)} \in \{-1, 1\}, \text{ so } |y^{(i)}| = 1 \\ &\leq \sum_{j \neq i} |e^{\frac{-\epsilon^2}{\tau^2}}| && ||x^{(j)} - x^{(i)}|| \geq \epsilon \text{ when } i \neq j \\ &= (m - 1) e^{\frac{-\epsilon^2}{\tau^2}} \end{aligned}$$

$$
\begin{aligned}
\text{Now we set this} &< 1 \\
(m-1)e^{\frac{-\epsilon^2}{\tau^2}} &< 1 \\
e^{\frac{-\epsilon^2}{\tau^2}} &< \frac{1}{m-1} \\
\log e^{\frac{-\epsilon^2}{\tau^2}} &< \log \frac{1}{m-1} \\
\frac{-\epsilon^2}{\tau^2} &< \log \frac{1}{m-1} \\
\frac{\epsilon^2}{\tau^2} &> -\log \frac{1}{m-1} \\
\frac{\epsilon^2}{\tau^2} &> \log(m-1) \\
\frac{\epsilon^2}{\log(m-1)} &> \tau^2 \\
\sqrt{\frac{\epsilon^2}{\log(m-1)}} &> \tau \\
\tau &< \frac{\epsilon}{\sqrt{\log(m-1)}}
\end{aligned}
$$

b)

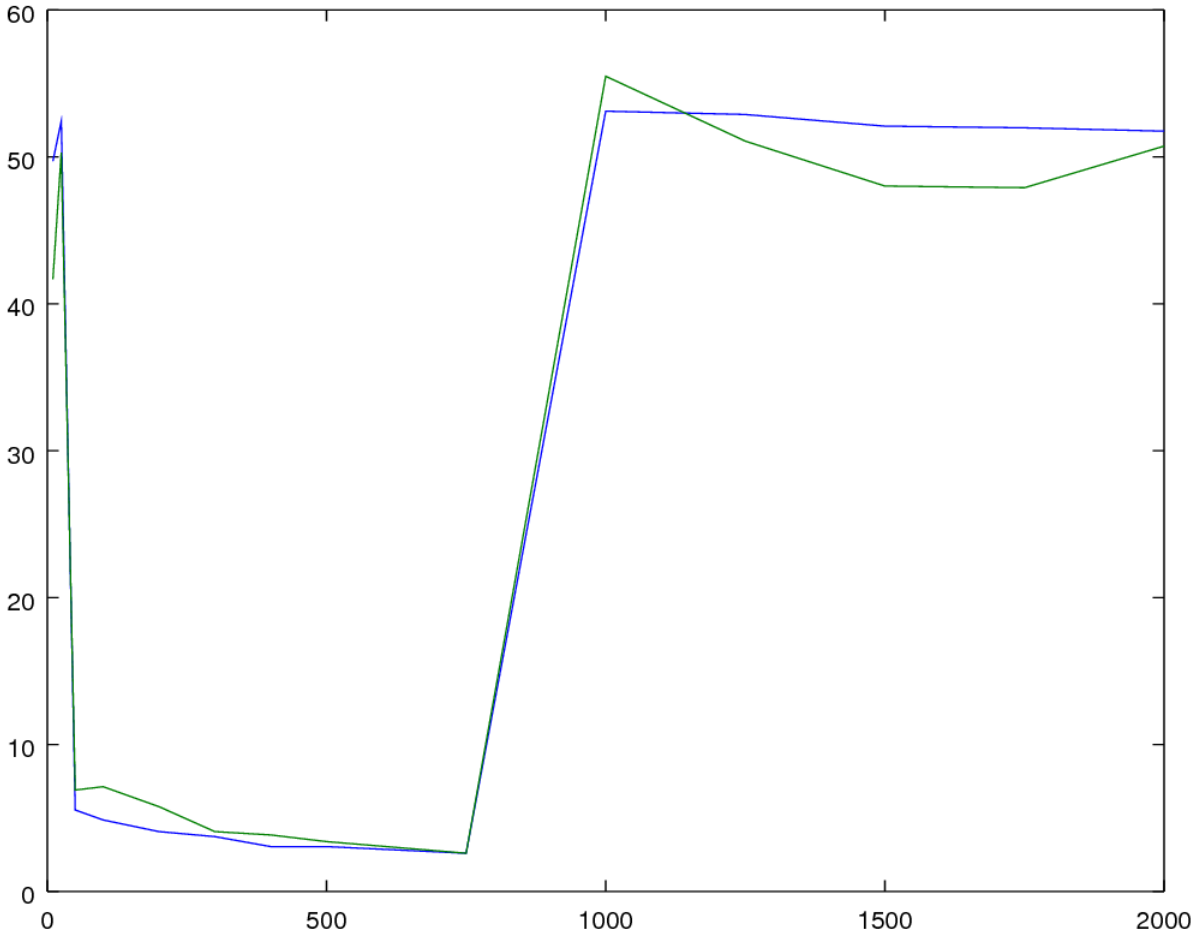
Yes, if we run an SVM without slack variables using the parameter τ from part (a), the resulting classifier will obtain zero training error. Zero training error corresponds to perfectly separating the training data, which is what an SVM without slack variables will do. Since our value of τ guarantees the training data to be separable, our classifier will obtain zero training error.

c)

No, if we run the SMO algorithm with slack variables the resulting classifier will not necessarily obtain zero training error. The parameter C controls how much we penalize the slack terms (ξ_i) and if we don't penalize slack enough then a smaller value for the objective function may be achieved by introducing a lot of slack and not necessarily by achieving zero training error.

4

Below is the graph I produced running weka against all the training sets. The x axis is number of training samples, while the y axis is percentage of incorrectly classified instances. The blue line is the NaiveBayesMultinomial classifier while the green line is the SMO classifier.



The result I expected is that both algorithms would improve as the number of training examples increased, with perhaps one outperforming the other. As you can see, something is likely wrong with the training data or with how I am running weka against it.

5

a)

We know that
$\exists h^* \in \mathcal{H} \quad . \quad \varepsilon(h^*) = 0$

Given $\hat{h} \in \mathcal{H} \quad . \quad \hat{\varepsilon}(\hat{h}) = 0$, prove that with probability $1 - \delta$, $\varepsilon(\hat{h}) \leq \frac{1}{m} \log \frac{k}{\delta}$

Consider a hypothesis $h_i \in \mathcal{H}$ such that $\varepsilon(h_i) > \gamma$
Let (x, y) be a sample drawn from the same distribution as the training data,
and let Z be a random variable where $Z = 1\{h_i(x) \neq y\}$

Then $P(Z) > \gamma$ and $P(\neg Z) \leq 1 - \gamma$

Let Z_j be the value of random variable Z evaluated against the j^{th} training example, i.e. $Z_j = 1\{h_i(x^{(j)}) \neq y^{(j)}\}$

Then the probability that our hypothesis h_i makes no mistakes on the training data is:
$P\big((\varepsilon(h_i) > \gamma) \wedge (\neg Z_1 \wedge \neg Z_2 \wedge ... \wedge \neg Z_m)\big) \leq (1-\gamma)^m$

No mistakes on the training data equates to zero training error, therefore:
$P\big((\varepsilon(h_i) > \gamma) \wedge (\hat{\varepsilon}(h_i) = 0)\big) \leq (1-\gamma)^m$

And the probability that the above holds for any one hypothesis in $\mathcal{H}$ is:

$$\begin{aligned} P\big(\exists h_i \in \mathcal{H}.(\varepsilon(h_i) > \gamma) \wedge (\hat{\varepsilon}(h_i) = 0)\big) &\leq k(1-\gamma)^m && \text{(by the union bound)} \\ &\leq ke^{-\gamma m} && \text{(by the inequality given in the problem)} \end{aligned}$$

Let $\delta = ke^{-\gamma m}$. Then

$$\begin{aligned} \frac{\delta}{k} &= e^{-\gamma m} \\ \log\frac{\delta}{k} &= -\gamma m \\ \gamma &= -\frac{1}{m}\log\frac{\delta}{k} \\ \gamma &= \frac{1}{m}\log\frac{k}{\delta} \end{aligned}$$

Then

$$\begin{aligned} P\big(\exists h_i \in \mathcal{H}.(\varepsilon(h_i) > \gamma) \wedge (\hat{\varepsilon}(h_i) = 0)\big) &\leq \delta && \text{(by the definition of } \delta \text{ above)} \\ P\big(\exists h_i \in \mathcal{H}.(\varepsilon(h_i) > \frac{1}{m}\log\frac{k}{\delta}) \wedge (\hat{\varepsilon}(h_i) = 0)\big) &\leq \delta && \text{(substitute in the value of } \gamma) \\ P\big(\neg\exists h_i \in \mathcal{H}.(\varepsilon(h_i) > \frac{1}{m}\log\frac{k}{\delta}) \wedge (\hat{\varepsilon}(h_i) = 0)\big) &> 1-\delta && \text{(subtract both sides from 1)} \\ P\big(\forall h_i \in \mathcal{H}.(\varepsilon(h_i) \leq \frac{1}{m}\log\frac{k}{\delta}) \vee (\hat{\varepsilon}(h_i) \neq 0)\big) &> 1-\delta \end{aligned}$$

Since $\hat{h} \in \mathcal{H}$, then with probability $> 1-\delta$,
$(\varepsilon(\hat{h}) \leq \frac{1}{m}\log\frac{k}{\delta}) \vee (\hat{\varepsilon}(\hat{h}) \neq 0)$

And since we know that $\hat{\varepsilon}(\hat{h}) = 0$, then with probability $> 1-\delta$,
$\varepsilon(\hat{h}) \leq \frac{1}{m}\log\frac{k}{\delta}$

b)

From part (a) we know that
$\gamma = \frac{1}{m}\log\frac{k}{\delta}$

Solving for m we get
$m = \frac{1}{\gamma}\log\frac{k}{\delta}$

So for fixed δ and γ, in order for $\varepsilon(\hat{h}) \leq \gamma$ to hold with probability $\geq 1-\delta$, it suffices that
$m \geq \frac{1}{\gamma}\log\frac{k}{\delta}$