

Stanford CS 229, Public Course, Problem Set 3

Dylan Price

December 26, 2016

1

a)

By the Hoeffding inequality, we know that

$$P(|\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| > \gamma) \leq 2 \exp(-2\gamma^2 \beta m)$$

Let A_i denote the event that $|\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| > \gamma$. Then

$$\begin{aligned} P(\exists \hat{h}_i \in \{\hat{h}_1 \dots \hat{h}_k\} . |\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| > \gamma) &= P(A_1 \cup \dots \cup A_k) \\ &\leq \sum_k P(A_i) \\ &\leq \sum_k 2 \exp(-2\gamma^2 \beta m) \\ &= 2k \exp(-2\gamma^2 \beta m) \end{aligned}$$

Therefore,

$$\begin{aligned} P(\neg \exists \hat{h}_i \in \{\hat{h}_1 \dots \hat{h}_k\} . |\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| > \gamma) \\ &= P(\forall \hat{h}_i \in \{\hat{h}_1 \dots \hat{h}_k\} . |\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| \leq \gamma) \\ &\geq 1 - 2k \exp(-2\gamma^2 \beta m) \end{aligned}$$

Let $\frac{\delta}{2} = 2k \exp(-2\gamma^2 \beta m)$. Then

$$\begin{aligned} \frac{\delta}{4k} &= \exp(-2\gamma^2 \beta m) \\ \frac{4k}{\delta} &= \exp(2\gamma^2 \beta m) \\ \log \frac{4k}{\delta} &= 2\gamma^2 \beta m \\ \frac{1}{2\beta m} \log \frac{4k}{\delta} &= \gamma^2 \\ \gamma &= \sqrt{\frac{1}{2\beta m} \log \frac{4k}{\delta}} \end{aligned}$$

Therefore,

$$P\left(\forall \hat{h}_i \in \{\hat{h}_1 \dots \hat{h}_k\} . |\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| \leq \sqrt{\frac{1}{2\beta m} \log \frac{4k}{\delta}}\right) \geq 1 - \frac{\delta}{2}$$

b)

From part (a), we have that

$$P(|\varepsilon(\hat{h}_i) - \hat{\varepsilon}_{S_{cv}}(\hat{h}_i)| \leq \gamma) \geq 1 - \frac{\delta}{2}, \text{ where } \gamma = \sqrt{\frac{1}{2\beta m} \log \frac{4k}{\delta}}$$

Because $\hat{h} \in \{\hat{h}_1 \dots \hat{h}_k\}$

$$P(|\varepsilon(\hat{h}) - \hat{\varepsilon}_{S_{cv}}(\hat{h})| \leq \gamma) \geq 1 - \frac{\delta}{2}$$

Let $h^* = \arg \min_{\hat{h}_i \in \{\hat{h}_1 \dots \hat{h}_k\}} \varepsilon(\hat{h}_i)$

Then with probability at least $1 - \frac{\delta}{2}$

$$\varepsilon(\hat{h}) \leq \hat{\varepsilon}_{S_{cv}}(\hat{h}) + \gamma$$

$$\leq \hat{\varepsilon}_{S_{cv}}(h^*) + \gamma$$

$$\leq \varepsilon(h^*) + 2\gamma$$

$$= \min_{i=1, \dots, k} \varepsilon(\hat{h}_i) + 2\gamma$$

$$= \min_{i=1, \dots, k} \varepsilon(\hat{h}_i) + 2\sqrt{\frac{1}{2\beta m} \log \frac{4k}{\delta}}$$

$$= \min_{i=1, \dots, k} \varepsilon(\hat{h}_i) + \sqrt{\frac{2}{\beta m} \log \frac{4k}{\delta}}$$

By the definition of \hat{h} it has the lowest $\hat{\varepsilon}_{S_{cv}}$ of any $\hat{h}_i \in \{\hat{h}_1 \dots \hat{h}_k\}$

By the uniform convergence result proved in part (a)

By the definition of h^*

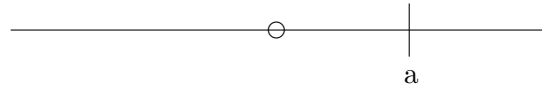
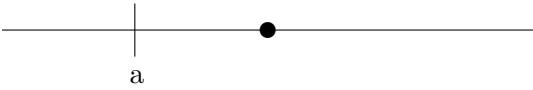
c)

2

a)

$$h(x) = 1\{a < x\}, \quad a \in \mathbb{R}$$

$h(x)$ can shatter a set of 1 point:



There is no set of 2 points that $h(x)$ can shatter because in the labeling situation shown below, no choice of a will successfully label both the points:



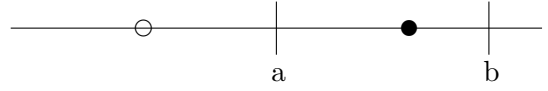
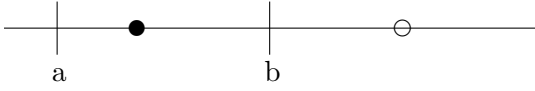
Therefore $VC(h(x)) = 1$

b)

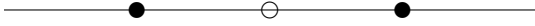
$$h(x) = 1\{a < x < b\}, \quad a, b \in \mathbb{R}$$

$h(x)$ can shatter a set of 2 points:





There is no set of 3 points that $h(x)$ can shatter because in the labeling situation shown below, no choice of a and b will successfully label all the points:



Therefore, $VC(h(x)) = 2$

c)

$$h(x) = 1\{a \sin x > 0\}, \quad a \in \mathbb{R}$$

$h(x)$ can shatter a set of 1 point by manipulating the sign of a . This will ensure that $h(x)$ can evaluate to a 1 or a 0 no matter where the point lies.

There is no set of 2 points that $h(x)$ can shatter. $h(x)$ divides the input space into π -wide sections which alternate evaluating to 1 or 0. E.g. with $a = 1$, $x \in (0, \pi)$ will evaluate to 1, $x \in [\pi, 2\pi]$ will evaluate to 0, $x \in (2\pi, 3\pi)$ will evaluate to 1, etc. Because of this, with any set of two points there is a labeling which $h(x)$ cannot achieve. There are two cases to consider: the two points both lie within sections that evaluate the same (i.e. both 1 or both 0) or they lie within sections that evaluate differently. In the first case, if the points are labeled differently $h(x)$ will not be able to achieve the labeling because no matter how you manipulate a both points will evaluate to 1 or both points will evaluate to 0. Similarly in the second case, if the points are labeled the same $h(x)$ will not be able to achieve the labeling because no matter how you manipulate a one point will evaluate to 1 and the other point will evaluate to 0.

Therefore $VC(h(x)) = 1$

d)

$$h(x) = 1\{\sin(x + 1) > 0\}, \quad a \in \mathbb{R}$$

$h(x)$ can shatter a set of 1 point:

$$\begin{aligned} x = \{0\} \quad \text{label } \{1\}, \quad a = \frac{\pi}{2} \\ \text{label } \{0\}, \quad a = 0 \end{aligned}$$

$h(x)$ can shatter a set of 2 points:

$$\begin{aligned} x = \{0, \frac{\pi}{2}\} \quad \text{label } \{0, 0\}, \quad a = -\frac{\pi}{2} \\ \text{label } \{0, 1\}, \quad a = 0 \\ \text{label } \{1, 0\}, \quad a = \frac{\pi}{2} \\ \text{label } \{1, 1\}, \quad a = \frac{\pi}{4} \end{aligned}$$

There is no set of 3 points which $h(x)$ can shatter because the points must span a distance of π to realize the labeling $\{0, 1, 0\}$, but in that case they cannot realize the labeling $\{1, 1, 1\}$.

Therefore $VC(h(x)) = 2$

3

a)

$$\begin{aligned}
J(\theta) &= \frac{1}{2} \|X\bar{\theta} + X_i\theta_i - \vec{y}\|_2^2 + \lambda \|\bar{\theta}\|_1 + \lambda s_i \theta_i \\
&= \frac{1}{2} (X\bar{\theta} + X_i\theta_i - \vec{y})^T (X\bar{\theta} + X_i\theta_i - \vec{y}) + \lambda \|\bar{\theta}\|_1 + \lambda s_i \theta_i \\
&= \frac{1}{2} (\bar{\theta}^T X^T + X_i^T \theta_i - \vec{y}^T) (X\bar{\theta} + X_i\theta_i - \vec{y}) + \lambda \|\bar{\theta}\|_1 + \lambda s_i \theta_i \\
&= \frac{1}{2} (\bar{\theta}^T X^T X \bar{\theta} + \bar{\theta}^T X^T X_i \theta_i - \bar{\theta}^T X^T \vec{y} + X_i^T \theta_i X \bar{\theta} + X_i^T \theta_i X_i \theta_i - X_i^T \theta_i \vec{y} - \vec{y}^T X \bar{\theta} - \vec{y}^T X_i \theta_i + \vec{y}^T \vec{y}) + \lambda \|\bar{\theta}\|_1 + \lambda s_i \theta_i
\end{aligned}$$

Find $\frac{\partial}{\partial \theta_i} J(\theta)$ when $s_i = 1$:

$$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{1}{2} (\bar{\theta}^T X^T X_i + X_i^T X \bar{\theta} + 2X_i^T X_i \theta_i - 2X_i^T \vec{y}) + \lambda$$

Set equal to 0:

$$\begin{aligned}
0 &= \frac{1}{2} (\bar{\theta}^T X^T X_i + X_i^T X \bar{\theta} + 2X_i^T X_i \theta_i - 2X_i^T \vec{y}) + \lambda \\
-X_i^T X_i \theta_i &= \frac{1}{2} \bar{\theta}^T X^T X_i + \frac{1}{2} X_i^T X \bar{\theta} - X_i^T \vec{y} + \lambda \\
\theta_i &= \frac{-1}{X_i^T X_i} \left(\frac{1}{2} \bar{\theta}^T X^T X_i + \frac{1}{2} X_i^T X \bar{\theta} - X_i^T \vec{y} + \lambda \right)
\end{aligned}$$

When $s_i = 1$, $\theta_i = \max \left(0, \frac{-1}{X_i^T X_i} \left(\frac{1}{2} \bar{\theta}^T X^T X_i + \frac{1}{2} X_i^T X \bar{\theta} - X_i^T \vec{y} + \lambda \right) \right)$

Following the same process for $s_i = -1$, we find

when $s_i = -1$, $\theta_i = \min \left(0, \frac{-1}{X_i^T X_i} \left(\frac{1}{2} \bar{\theta}^T X^T X_i + \frac{1}{2} X_i^T X \bar{\theta} - X_i^T \vec{y} - \lambda \right) \right)$

b)

See q3/l1ls.m

c)

Following are the thetas arrived at by the coordinate descent algorithm from part (b), vertical dots indicate areas of contiguous 0s.

$$\lambda = .001, \quad [\text{index}, \theta] = \begin{bmatrix} 1 & 0.67077 \\ 2 & 0.81233 \\ 3 & -0.82544 \\ 4 & -0.81632 \\ 5 & -0.93817 \\ 6 & -0.00437 \\ & \vdots \\ 12 & -0.00632 \\ 13 & -0.00213 \\ 14 & 0.00000 \\ 15 & -0.03042 \\ & \vdots \\ 33 & 0.00349 \\ 34 & -0.01144 \\ & \vdots \\ 39 & 0.00203 \\ & \vdots \\ 48 & 0.01879 \\ & \vdots \\ 51 & 0.00512 \\ & \vdots \\ 71 & 0.00002 \\ 72 & 0.00000 \\ 73 & 0.00033 \\ 74 & -0.02040 \\ & \vdots \\ 77 & 0.00023 \\ & \vdots \\ 81 & 0.00059 \\ 82 & 0.01227 \\ 83 & -0.01478 \\ & \vdots \\ 86 & -0.00727 \\ & \vdots \\ 91 & 0.00026 \\ & \vdots \\ 94 & -0.00046 \\ & \vdots \end{bmatrix}$$

$$\lambda = .01, \quad [\text{index}, \theta] = \begin{bmatrix} 1 & 0.66890 \\ 2 & 0.81693 \\ 3 & -0.81803 \\ 4 & -0.81682 \\ 5 & -0.94517 \\ 6 & -0.00377 \\ & \vdots \\ 12 & -0.00393 \\ 13 & -0.00015 \\ 14 & 0.00000 \\ 15 & -0.03132 \\ & \vdots \\ 34 & -0.00887 \\ & \vdots \\ 39 & 0.00176 \\ & \vdots \\ 48 & 0.01794 \\ & \vdots \\ 71 & 0.00093 \\ & \vdots \\ 74 & -0.02258 \\ & \vdots \\ 81 & 0.00188 \\ 82 & 0.01066 \\ 83 & -0.01606 \\ & \vdots \\ 86 & -0.00284 \\ 87 & 0.00000 \\ & \vdots \\ 95 & 0.00194 \\ & \vdots \end{bmatrix}$$

$$\lambda = .1, \quad [\text{index}, \theta] = \begin{bmatrix} 1 & 0.65289 \\ 2 & 0.81030 \\ 3 & -0.80756 \\ 4 & -0.80795 \\ 5 & -0.95211 \\ & \vdots \\ 12 & -0.00537 \\ 13 & -0.00216 \\ 14 & 0.00000 \\ 15 & -0.02394 \\ & \vdots \\ 18 & 0.00292 \\ & \vdots \\ 34 & -0.00294 \\ & \vdots \\ 48 & 0.00484 \\ & \vdots \\ 74 & -0.01429 \\ & \vdots \\ 83 & -0.00298 \\ & \vdots \\ 86 & -0.00261 \\ & \vdots \end{bmatrix} \quad \lambda = 1, \quad \theta = \begin{bmatrix} 0.49876 \\ 0.65611 \\ -0.79058 \\ -0.65564 \\ -0.89192 \\ \vdots \end{bmatrix}$$

$$\lambda = 10, \quad \theta = \begin{bmatrix} 0.00000 \\ 0.00000 \\ -0.41080 \\ 0.00000 \\ -0.07083 \\ \vdots \end{bmatrix}$$

For comparison,

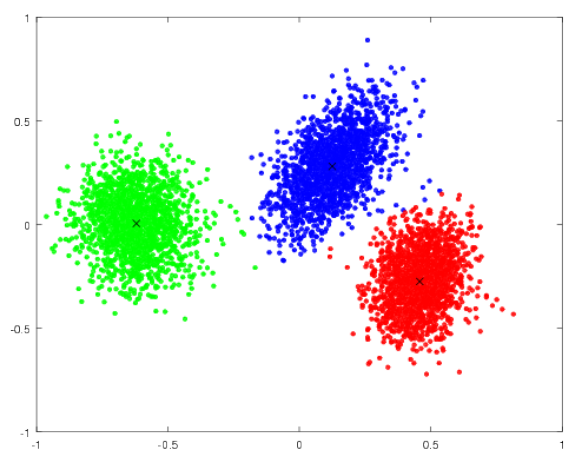
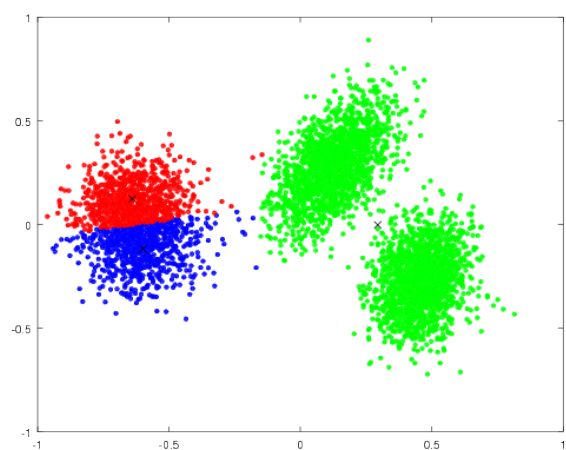
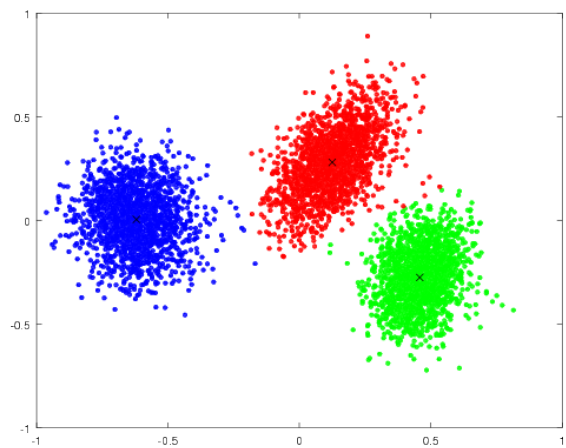
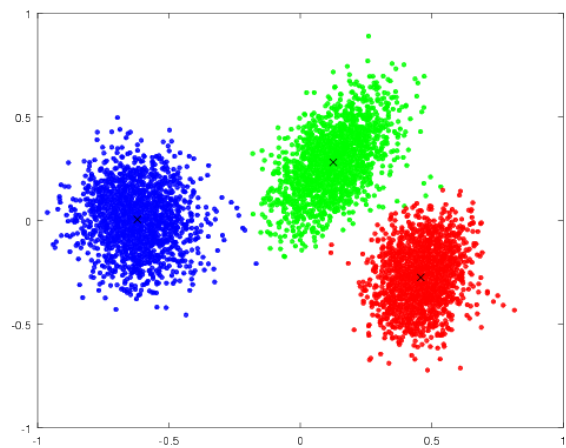
$$\theta_{\text{true}} = \begin{bmatrix} 0.68372018 \\ 0.84110202 \\ -0.83028605 \\ -0.85031124 \\ -0.93904984 \\ \vdots \end{bmatrix}$$

As λ is increased, the θ output by the algorithm becomes more sparse. This algorithm could be used for feature selection because most of the values in θ are set to zero, meaning those features are effectively ignored. By tuning λ you can control approximately how many features are kept by the algorithm.

4

See q4/k_means.m for code

$k = 3$



$k = 4$

