Quiz 1
October 26, 2015

Time: 50 minutes; Total Marks: 38
One double-sided 8.5” x 11” cheat sheet allowed

This test contains 4 questions and 6 pages

NAME:

STUDENT NUMBER:

Question	Marks	Time budget
1	/16	15 min
2	/6	10 min
3	/10	10 min
4	/6	10 min

might move the regressor away from the test data

1. (16 marks) True or False questions. **No explanation required.**

   (a) True or $\boxed{\text{False.}}$ Test error always decreases when more training data are used.

   you may add some samples that are different from test data
   new sample might be noisy as well. For example, in case of regression, new sample

   (b) True or False. When modeling coin tossing, the maximum a posteriori estimate for μ is the same as the maximum likelihood estimate if a "flat" prior is used:

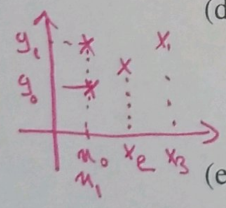
   Posterior $\propto$ prior $\times$ likelihood

   $$p(\mu) = \begin{cases} 1 & 0 \le \mu \le 1 \\ 0 & o.w. \end{cases}$$

   Flat → multiplying it has no effect → maximum a posteriori estimate is the same as the maximum likelihood estimate

   (c) True or False. $p(\boldsymbol{x}) \le 1$ for a Gaussian kernel density estimate:

   $$p(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} \exp\left\{ -\frac{||\boldsymbol{x} - \boldsymbol{x}_n||^2}{2h^2} \right\}$$



   (d) True or False. Given any fixed test set for regression, there always exists a set of polynomials that gives zero error on this test set.

   ~~the dataset have not~~ datasets might have noisy sample. For example, you might have two samples with the same set of features but different labels.
   False

   (e) True or False. When training logistic regression with gradient descent, each iteration of gradient descent will cause the error (negative log likelihood) to decrease.

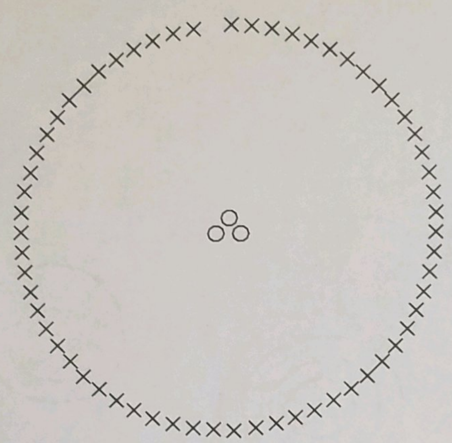
   If step-size is too large → it can increase the error

   (f) True or False. Kernelized perceptron can produce non-linear decision boundaries in the original input space.

   (g) True or False. If $k_1(\boldsymbol{x}, \boldsymbol{z})$ is a valid kernel, then $k_2(\boldsymbol{x}, \boldsymbol{z}) = k_1(\boldsymbol{x}, \boldsymbol{z}) + 1$ is always valid too.

   (h) True or False. Removing a training data point which is a support vector will cause the SVM decision boundary to move.

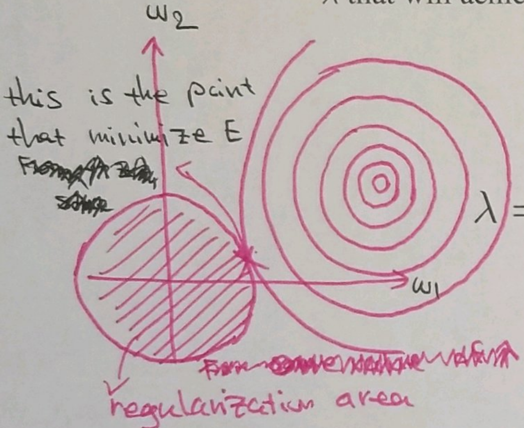
2. (6 marks) Consider using a **K nearest neighbour** classification model with the training set shown below. Suppose we use leave-one-out cross-validation (LOO-CV) to determine the value of the parameter K from $\{1, 3, 5, 7, \ldots\}$. Explain what the result of this procedure would be.

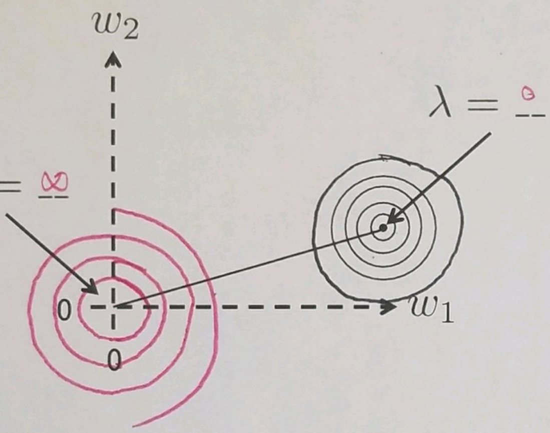


3. (10 marks) Recall regularized regression:

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

- (3 marks) The picture below shows the minimum of squared error and isocurves of equal squared error. Label the ends of the solid line segment according to the values of λ that will achieve these values for parameters $\boldsymbol{w}$.

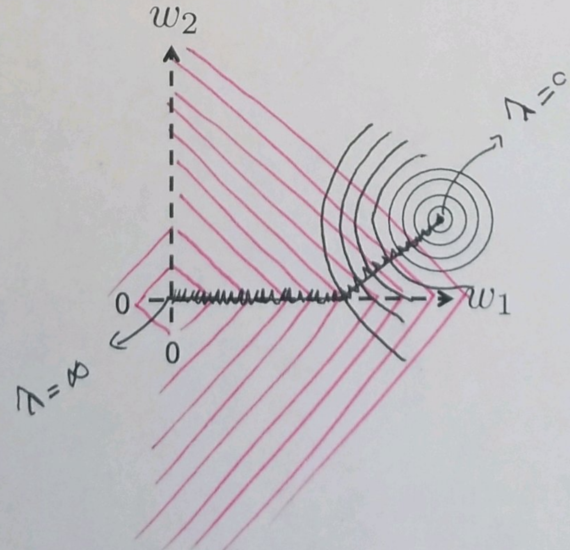




$||w||^2$

$w_0^2 + w_1^2 \leq \xi$

- (3 marks) Draw a similar picture for L_1 regularization (lasso). Draw the equivalent to the solid line and label its ends.



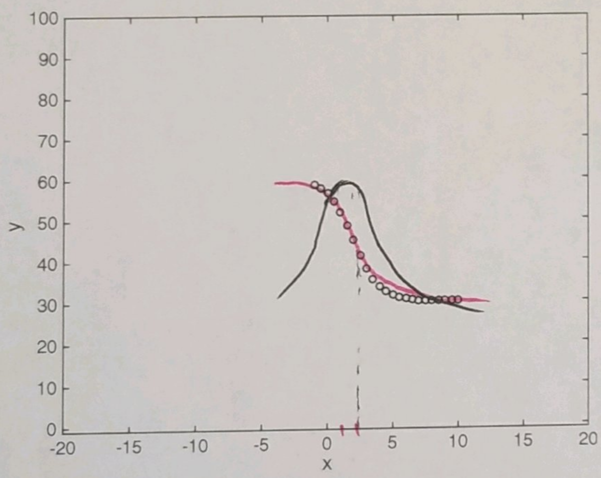
$|w_0| + |w_1| \leq \xi$

- (4 marks) Consider Gaussian versus sigmoid basis functions for un-regularized regression on the 1-d dataset below. Draw 2 curves: from using (a) $\phi_g(x)$ and a bias term; or (b) $\phi_s(x)$ and a bias term.

$$\phi_g(x) = \exp\left\{-\frac{(x-1)^2}{4}\right\} \rightarrow \text{Gaussian}$$

$$y(x) = w_0 + w_1 \phi(x)$$

$$\phi_s(x) = \frac{1}{1+\exp(2-x)} \rightarrow \text{sigmoid}$$



4. (6 marks) Consider training a support vector machine with a linear kernel on a **linearly separable dataset**. Is there any difference in the hyperplane $(\boldsymbol{w}, b)$ found using the exact (hard margin) classification constraints:

$$\begin{aligned} & \arg\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \\ s.t. \quad & \forall n\,,\, t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) \geq 1 \end{aligned}$$

and using those with slack variables (soft margin):

$$\begin{aligned} & \arg\min_{\boldsymbol{w},b,\xi_n} C\sum_{n=1}^{N} \xi_n + \frac{1}{2}||\boldsymbol{w}||^2 \\ s.t. \quad & \forall n\,,\, t_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) \geq 1 - \xi_n \\ & \forall n\,,\, \xi_n \geq 0 \end{aligned}$$

State whether there is a difference for a **linearly separable dataset**. If so, explain and show an example of the different behaviour. If not, give a brief argument why not.