NAME:_____ **SID#:**_____ **Section:**_____

CS 194-10 Fall 2011 Introduction to Machine Learning Stuart Russell Midterm

You have 80 minutes. The exam is open-book (class-designated reading materials only), open-notes. 80 points total. Panic not.

Mark your answers ON THE EXAM ITSELF. Write your name, SID, and section number at the top of each sheet. For true/false questions, CIRCLE *True* OR *False*.
For multiple-choice questions, CIRCLE *ALL* CORRECT CHOICES (in some cases, there may be more than one).

If you are not sure of your answer you can provide a *brief* explanation; we will try to give partial credit where appropriate.

For official use only

Q. 1	Q. 2	Q. 3	Q. 4	Q. 5	Total
/20	/10	/10	/20	/20	/80

1. (20 pts.) Some Easy Questions to Start With

(a) (4) *True/False*: In a least-squares linear regression problem, adding an L_2 regularization penalty cannot decrease the L_2 error of the solution $\hat{\mathbf{w}}$ on the training data.

(b) (4) *True/False*: In a least-squares linear regression problem, adding an L_2 regularization penalty always decreases the expected L_2 error of the solution $\hat{\mathbf{w}}$ on unseen test data.

(c) (4) *True/False*: In a regression problem, decision trees with constant leaves can fit every data set with zero training error.

(d) (4) *True/False*: As the width b goes to 0 in locally weighted regression with a quadratic kernel $K(d) = max(0, 1 - d^2/b^2)$, the algorithm behaves exactly like 1-nearest-neighbor.

(e) (4) *True/False*: In decision tree learning with noise-free data, starting with the wrong attribute at the root can make it impossible to find a tree that fits the data exactly.

2. (10 pts.) Locally weighted regression
Can locally weighted regression exactly reproduce the learning behavior of ordinary least-squares regression, given a suitable kernel, for any data set? If so, how? If not, why not?

3. (10 pts.) Regularization
The standard form of the L_2-regularized L_2 loss function for linear regression is

$$L = (\mathbf{Y} - \mathbf{X}\mathbf{w})^T(\mathbf{Y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T\mathbf{w} \quad \text{where } \lambda > 0.$$

(a) (3) Suppose we accidentally write $L = (\mathbf{Y} - \mathbf{X}\mathbf{w})^T(\mathbf{Y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{Y}^T\mathbf{Y}$ instead. Explain why this form of "regularization" has no effect.

(b) (3) Suppose we use the correct expression but accidentally choose $\lambda < 0$. Explain briefly how this defeats the purpose of regularization.

(c) (4) In ordinary least squares, the squared-error loss measures error in the y direction. In *total least squares*, error is measured by the orthogonal distance from the point to the line (i.e., the length of the perpendicular). Explain briefly why regularizing with $\lambda < 0$ would be disastrous in this case. You may find it helpful to consider univariate regression ($y = w_0 + w_1 x$) as an example.

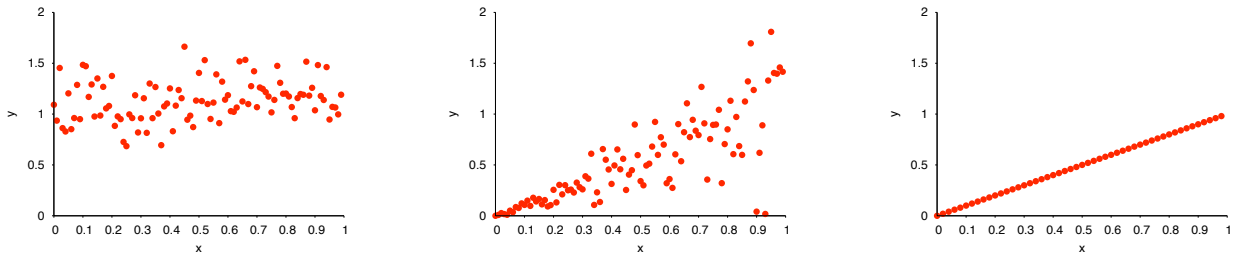

Figure 1: Three data sets.

4. (20 pts.) Input-dependent noise in regression

Ordinary least-squares regression is equivalent to assuming that each data point is generated according to a linear function of the input plus zero-mean, constant-variance Gaussian noise. In many systems, however, the noise variance is itself a positive linear function of the input (which is assumed to be non-negative, i.e., $x \geq 0$).

(a) (5) Which of the following families of probability models correctly describes this situation in the univariate case? (Hint: only one of them does.)

i.
$$P(y \mid x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(\frac{(y - (w_0 + w_1 x))^2}{2x^2\sigma^2}\right)$$

ii.
$$P(y \mid x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(\frac{(y - (w_0 + (w_1 + \sigma^2)x))^2}{2\sigma^2}\right)$$

iii.
$$P(y \mid x) = \frac{1}{\sigma \sqrt{2\pi x}} \exp\left(\frac{(y - (w_0 + w_1 x))^2}{x\sigma^2}\right)$$

(b) (6) Circle the plots in Figure 1 that could plausibly have been generated by some instance of the model family(ies) you chose.

(c) (3) *True/False*: Regression with input-dependent noise gives the same solution as ordinary regression for an infinite data set generated according to the corresponding model.

(d) (6) For the model you chose in part (a), write down the derivative of the negative log likelihood with respect to w_1.

5. (20 pts.) Classifying data

(i)

X_1	X_2	Y
1	1	+
4	2	−
4	5	−
5	5	+

(ii)

X_1	X_2	Y
1	1	+
5	5	−
4	5	−
5	5	+

(iii)

X_1	X_2	Y
1	1	+
4	2	−
4	5	+
5	5	+

(a) (3) *Multiple choice*: Which data sets are linearly separable?
(i) (ii) (iii)

(b) (3) *Multiple choice*: Which data sets have label noise?
(i) (ii) (iii)

(c) (3) *Multiple choice*: Which data sets can be fit exactly by a decision tree?
(i) (ii) (iii)

(d) (5) A *1-decision-list* is a decision tree in which the "yes" branch of every binary test is a leaf node. For a continuous attribute X_j, a test can be either $X_j > c$ or $X_j < c$. Continuous attributes can appear in multiple tests. Pick a data set and show a decision list that fits it exactly.

(e) (6) In the absence of label noise, can any two-class data set in two dimensions be fit exactly by a decision list? Briefly explain why, or give a counterexample.