

**Stat 342 Final Exam
Fall 2014**

I have neither given nor received unauthorized assistance on this exam.

Name Signed

Date

Name Printed

There are 16 questions on the following 9 pages. Do as many of them as you can in the available time. I will score each question out of 10 points AND TOTAL YOUR BEST 10 SCORES. (That is, this is a 100 point exam.)

1. Consider a continuous distribution for (y, x) on the unit square with joint pdf

$$f(y, x) = \frac{12}{13}(x + xy + y^2)I[0 < x < 1 \text{ and } 0 < y < 1]$$

10 pts a) Find the SEL predictor of y based on x , $\hat{y}^{\text{opt}}(x)$. (Give an explicit formula for this function.)

10 pts b) Suppose that one wishes to sample from the joint distribution specified by $f(y, x)$ above using a Gibbs (MCMC) algorithm. Suppose that in the process of generating a Gibbs sequence of (y, x) pairs, one has in hand $(y^{\text{current}}, x^{\text{current}})$. Give formulas for the conditional pdf's to be sampled:

if the next update replaces y^{current} . (Give a conditional pdf for the next y that depends upon x^{current} .)

if the next update replaces x^{current} . (Give a conditional pdf for the next x that depends upon y^{current} .)

10 pts c) Suppose that one wishes to sample from the joint distribution specified by $f(y, x)$ above using a rejection algorithm. Completely specify such an algorithm. (Describe completely **how you will generate** proposals (y^*, x^*) from a specific joint distribution specified by a joint pdf $g(y, x)$. Then say precisely/completely **how you will decide** whether or not to accept a proposal.)

10 pts 2. Suppose that one is furnished with $U \sim \text{Uniform}(0,1)$. Of interest is a function $h(u)$ so that $V = h(U)$ has pdf

$$f(v) = \left(\frac{1}{2} + v\right) I[0 < v < 1]$$

Find $h(u)$.

3. A particular distribution involves a parameter $\theta \in (0,1)$. The model is such that a single observation from the distribution carries Fisher information about θ that is

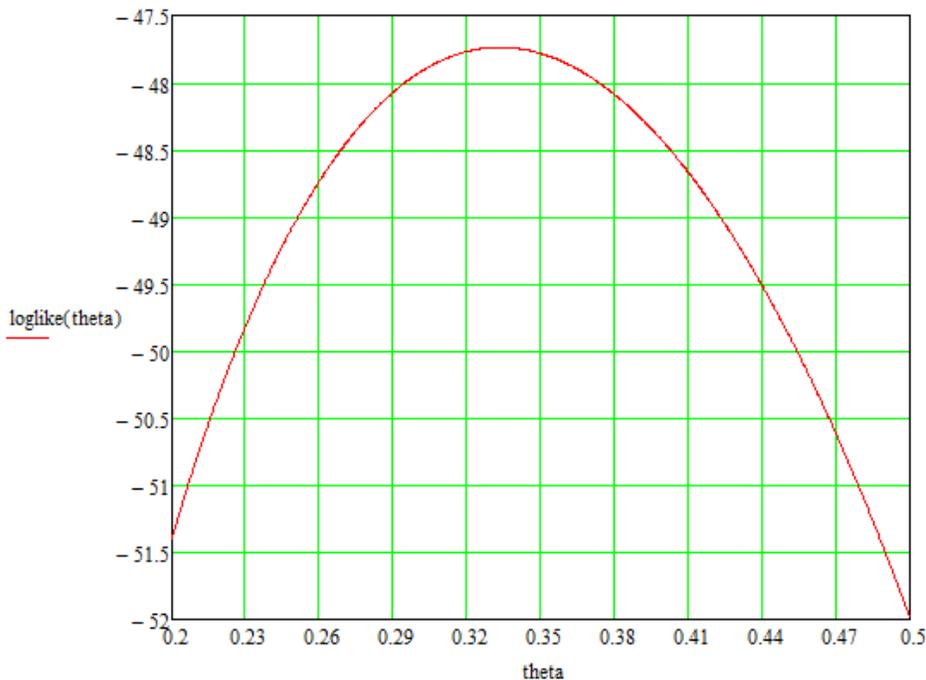
$$I(\theta) = \frac{2 - 3\theta + \theta^2}{\theta(1-\theta)}$$

10 pts a) Suppose that a sample of $N = 100$ iid observations from this distribution produces an MLE

$$\hat{\theta}^{\text{MLE}} = .5$$

Give approximately 95% confidence limits for θ based on this outcome.

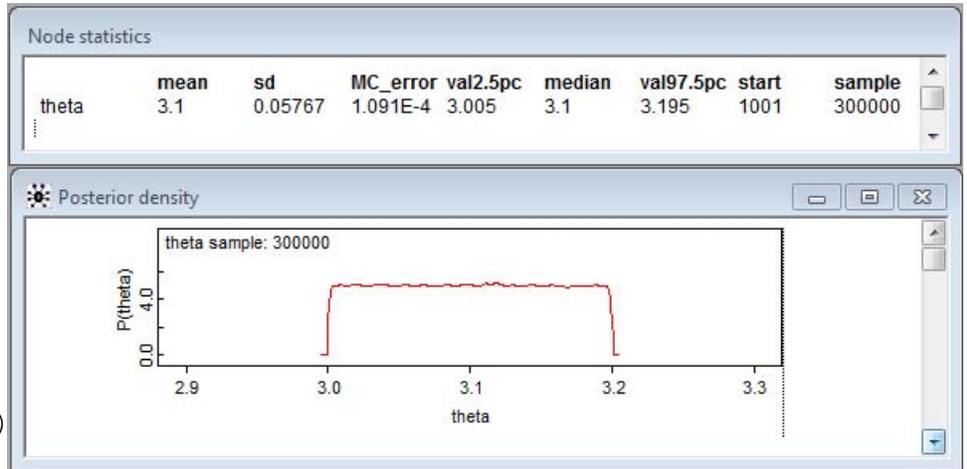
10 pts b) Another large iid sample from a statistical model produces a loglikelihood function as plotted below. What are $\hat{\theta}^{\text{MLE}}$ and approximate 95% two-sided confidence limits for θ based on this sample? (Indicate clearly how you obtain your answers.)



$\hat{\theta}^{\text{MLE}} =$ _____ 95% Confidence limits: _____ , _____

10 pts 4. Here is some BUGS code and output for a Bayes analysis in a statistical problem involving a parameter θ .

```
model {  
  theta ~ dunif(-10,10)  
  thetal <- theta - .5  
  thetau <- theta + .5  
  
  for (i in 1:5) {  
    x[i] ~ dunif(thetal,thetau)  
  }  
  
  #Here are the data  
  list(x=c(3.2,2.8,3.1,3.5,2.7))  
  #Here is a starting value  
  list(theta=3)
```



Here, what are the elements of the statistical model and inferences based on the data $x_1 = 3.2, x_2 = 2.8, x_3 = 3.1, x_4 = 3.5, x_5 = 2.7$? That is, identify the items indicated below.

The density for individual observations given the parameter θ , $f(x|\theta)$:

The prior density for θ , $g(\theta)$:

An approximate value for the SEL Bayes estimate of θ , $\hat{\theta}^{\text{opt}}(\text{data}) = E[\theta | \text{data}]$:

An approximate 95% credible interval for θ :

An approximate value for the posterior standard deviation of θ :

5. Below is a set of 3 distributions for a discrete statistical model given in tabular form. Use it in what follows.

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
$\theta = 3$.2	.2	.2	.3	.1
$\theta = 2$.2	.4	.1	.1	.2
$\theta = 1$.3	.2	.2	.2	.1

10 pts a) Identify any minimal sufficient statistic, $T(x)$, in this model. Give values below for this statistic.

$T(1) =$ _____

$T(2) =$ _____

$T(3) =$ _____

$T(4) =$ _____

$T(5) =$ _____

10 pts b) Consider a **3-class** classification problem with 0-1 loss. For a prior distribution with $g(1) = .2$, $g(2) = .3$, and $g(3) = .5$, find a Bayes optimal decision function $a^{\text{opt}}(x)$ and its risk function $R(\theta) = E_{\theta}L(\theta, a^{\text{opt}}(x))$.

$a^{\text{opt}}(1) =$ _____

$a^{\text{opt}}(2) =$ _____

$a^{\text{opt}}(3) =$ _____

$R(1) =$ _____

$a^{\text{opt}}(4) =$ _____

$R(2) =$ _____

$a^{\text{opt}}(5) =$ _____

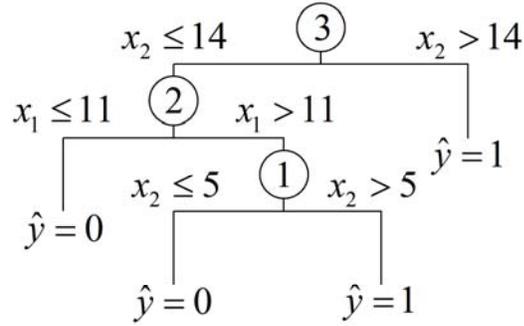
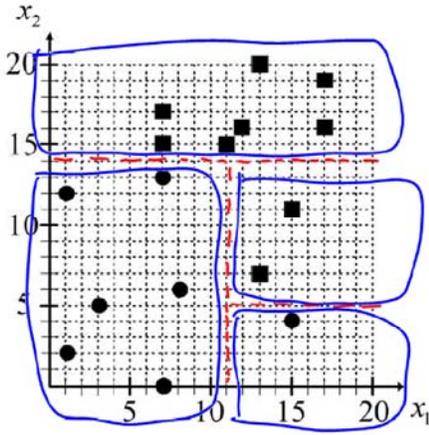
$R(3) =$ _____

10 pts

6. Below is a 2-class classification tree and graphic of corresponding rectangles for a $p = 2$ predictor problem based on a training sample of size $N = 16$. (This is the same toy problem that was used on Exam 2. Circles are $y = 0$ cases and squares are $y = 1$ cases.) Consider pruning this tree not at all, or at any of the points labeled (1), (2), and (3) (that is, consider 4 subtrees of this full tree). For each tree/subtree, find $\overline{\text{err}}$ and thus the corresponding function of α ,

$$C_\alpha = \# \text{ of rectangles} + \alpha \cdot \overline{\text{err}}$$

Then say what subtree is optimal for cost-complexity pruning with $\alpha = 4$.



Tree Pruned at

$\overline{\text{err}}$

None	
(1)	
(2)	
(3)	

The best tree/subtree for $\alpha = 4$ is the one pruned at: _____

10 pts

7. Write T (for true) or F (for false) before each of the following 5 statements.

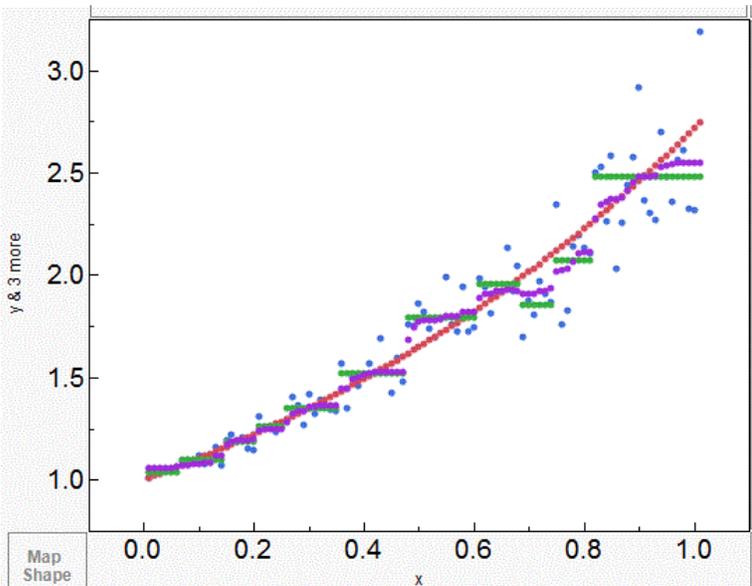
- _____ In general terms, larger data sets support the effective use of more complex predictors than do smaller data sets.
- _____ The number of bootstrap samples employed in a bagging predictor (B) is a complexity parameter for bagging.
- _____ Increasing a bandwidth parameter (λ) in a smoothing algorithm increases the complexity of the corresponding predictor.
- _____ Nearest neighbor and local regression smoothing predictors are particularly effective for high-dimensional (of the input vectors) prediction problems.
- _____ If one of a set of SEL predictors is essentially $E[y | \mathbf{x}]$, substantial improvement in that predictor can often be obtained through "ensemble" methods using the whole set.

- 10 pts** 8. Below is a small ($N = 4$) fake training set of pairs (y, x) (with centered y and standardized x). Find the ridge regression predictor of y for $\lambda = 2$, $\hat{y}_2^{\text{Ridge}}(x) = x \cdot \hat{\beta}_2^{\text{Ridge}}$ (that is, find $\hat{\beta}_2^{\text{Ridge}}$).

y	-4.0	-4.5	2.5	6.0
x	-1.0	-.7	.7	1.0

(Very simple 1-variable calculus can be used here.)

- 10 pts** 9. Below is a plot of a simple (x, y) training data set. Plotted also are the (smooth) conditional mean function and two sets of predictions made from the data. One set is from a single regression tree predictor and the other set is from a random forest predictor based on a fairly large number of trees. (Since the input is only one-dimensional there can be no random selection of features for each tree in the forest, and so the latter is really simply a bagged tree.)



In this picture, for x near 1.0, both sets of predictions are substantially below the conditional mean. Do you think this is purely a random phenomenon that would be corrected "on average" across many different training sets? If so, why? If not, why not?

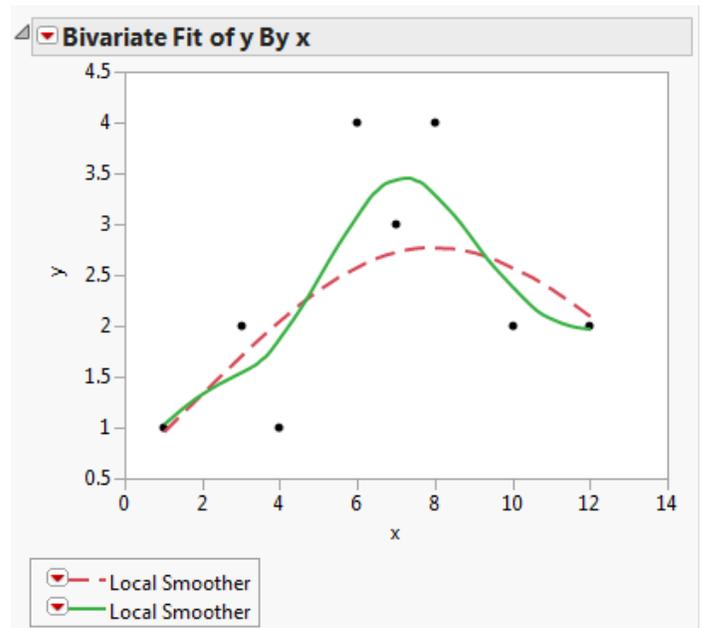
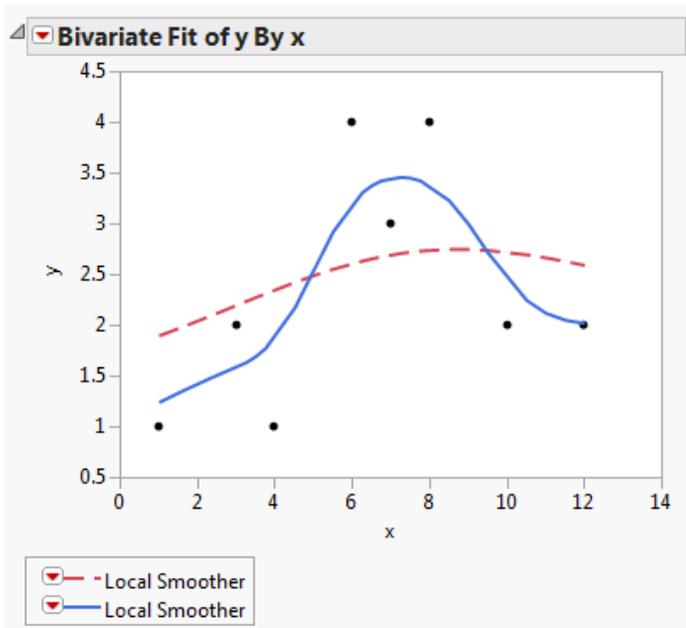
Is this likely just "random/the luck of the draw" for this training set?

YES or NO (**circle one**)

WHY?

10 pts

10. The JMP "Fit Y by X" procedure has a "Kernel Smoother" option. Below are graphics made using that option with both local "constant" and "linear" fitting and with both "small" and "large" bandwidth (on the toy example from class). Identify the plots according to degree (constant or linear) and bandwidth. (Circle the correct descriptor for each plot.)



Degree **Bandwidth**

Dashed: constant/linear small/large

Solid: constant /linear small/large

Degree **Bandwidth**

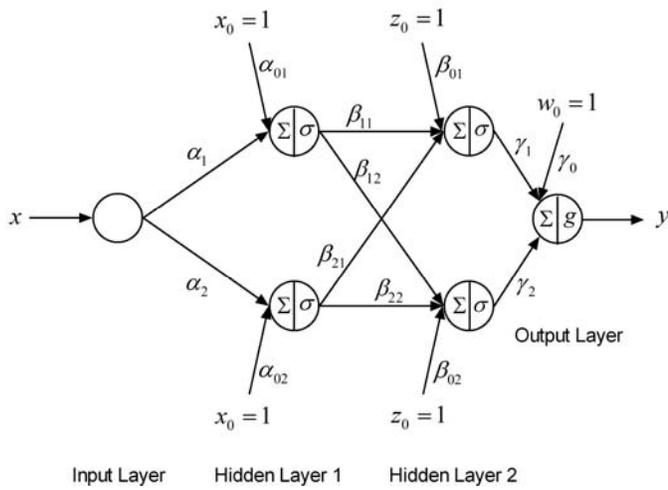
Dashed: constant/linear small/large

Solid: constant/linear small/large

Evaluate the Nadaraya-Watson smoother based on the standard normal kernel and bandwidth $\lambda = .5$ at the value $x = 7$.

10 pts

11. Below is a "network diagram" for a 2 hidden layer feedforward neural network. Suppose that an appropriate fitting method applied to a training set of (y, x) pairs produces the set of fitted coefficients next to the diagram.



$$\begin{aligned} \hat{\alpha}_{01} &= 5 & \hat{\beta}_{11} &= 2 & \hat{\gamma}_1 &= -505 \\ \hat{\alpha}_1 &= -2 & \hat{\beta}_{12} &= -2 & \hat{\gamma}_2 &= 850 \\ \hat{\alpha}_{02} &= 7 & \hat{\beta}_{21} &= -5 & \hat{\gamma}_0 &= 141 \\ \hat{\alpha}_2 &= -3 & \hat{\beta}_{22} &= 6 & & \\ & & \hat{\beta}_{01} &= 1 & & \\ & & \hat{\beta}_{02} &= -2 & & \end{aligned}$$

For "activation function" $\sigma(u) = 1/(1 + \exp(u))$ and $g(x) = x$, evaluate $\hat{y}^{\text{NeuralNet}}(2)$.

10 pts

12. Below is a cartoon of a hypothetical training 2-class classification training set of points (y, x_1, x_2) , presented in the same style as the diagram in problem 6. Suppose that a support vector classifier is of the form $I[x_1 + x_2 - 3 > 0]$. The "margin" of this classifier is then $M = 1/\|(1,1)\| = 1/\sqrt{2}$. On the plot indicate support vectors by circling them and misclassifications by x's. (Circles are $y = 0$ and squares are $y = 1$.)

