Stat 342 Final Exam
Fall 2014

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

KEY

_____

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

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

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

We need $E[y|x]$. This is

$$\frac{\frac{12}{13}\int_0^1 y\left(x + xy + y^2\right) dy}{\frac{12}{13}\int_0^1 \left(x + xy + y^2\right) dy} = \frac{x\left(\frac{1}{2}\right) + x\left(\frac{1}{3}\right) + \frac{1}{4}}{x + x\left(\frac{1}{2}\right) + \frac{1}{3}} = \frac{\left(\frac{5}{6}\right)x + \frac{1}{4}}{\left(\frac{3}{2}\right)x + \frac{1}{3}}$$

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 $\left(y^{\text{current}}, x^{\text{current}}\right)$. 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}.)

One samples a density proportional to

$$h(y) = x^c + x^c y + y^2$$

Since $\int_0^1 h(y)\,dy = x^{\text{current}} + \frac{1}{2}x^{\text{current}} + \frac{1}{3}$ this is

$$g\left(y \mid x^{\text{current}}\right) = \frac{x^{\text{current}} + x^{\text{current}} y + y^2}{\frac{3}{2}x^{\text{current}} + \frac{1}{3}}$$

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

One samples a density proportional to

$$h(x) = x + xy^c + (y^c)^2$$

Since $\int_0^1 h(x)\,dx = \frac{1}{2} + \frac{1}{2}y^{\text{current}} + \left(y^{\text{current}}\right)^2$ this is

$$g\left(x \mid y^{\text{current}}\right) = \frac{x + xy^{\text{current}} + \left(y^{\text{current}}\right)^2}{\frac{1}{2} + \frac{1}{2}y^{\text{current}} + \left(y^{\text{current}}\right)^2}$$

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.)

Use the uniform dsn on $[0,1]^2$ as a proposal distribution, i.e. with u_1^*, u_2^* iid $U(0,1)$ consider

$$\underset{\sim}{u}^* = (u_1^*, u_2^*)$$

as the proposal (This has density $I[\underset{\sim}{u}^* \in [0,1]^2] = g(\underset{\sim}{u}^*)$

Then $f(y,x) \le \frac{36}{13} g((x,y))$. So for U_3 independent of $\underset{\sim}{u}^*$ and $U(0,1)$, accept $\underset{\sim}{u}^*$ provided

$$\frac{36}{13} U_3 < f(u_1^*, u_2^*)$$

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 $h(u)$.

For F the desired cdf, we want $h(u) = F^{-1}(u)$.

For $v \in (0,1)$

$$F(v) = \int_0^v \left(\frac{1}{2} + x\right) dx = \frac{v}{2} + \frac{v^2}{2}$$

So if $F(v) = u$, $\frac{v^2}{2} + \frac{v}{2} - u = 0$

i.e. $\frac{-\frac{1}{2} \pm \sqrt{\frac{1}{4} + 2u}}{2(\frac{1}{2})}$ i.e. $v = -\frac{1}{2} + \sqrt{\frac{1}{4} + 2u}$

and $h(u) = -\frac{1}{2} + \sqrt{\frac{1}{4} + 2u}$

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

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

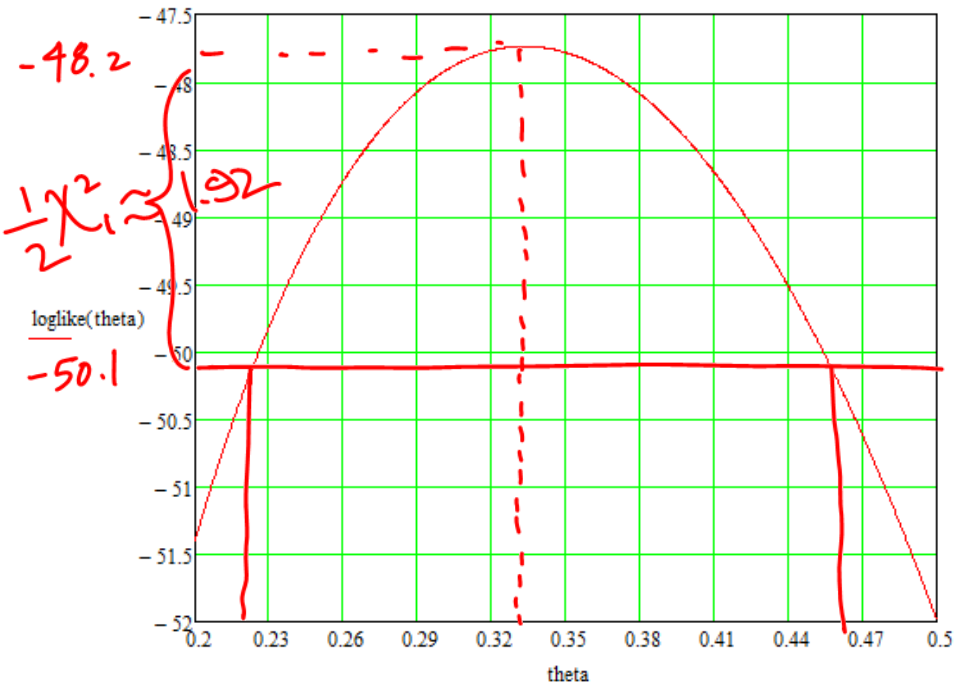
Give approximately 95% confidence limits for θ based on this outcome.

Use $\hat{\theta}^{\text{MLE}} \pm 2 \frac{1}{\sqrt{N I(\hat{\theta}^{\text{MLE}})}}$

i.e. $.5 \pm 1.96 \frac{1}{\sqrt{100\left(\frac{2-3(.5)+(.5)^2}{(.5)(1-.5)}\right)}}$

i.e. $.5 \pm .11$

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.)



$\hat{\theta}^{\text{MLE}} =$ .33 95% Confidence limits: .22 , .46

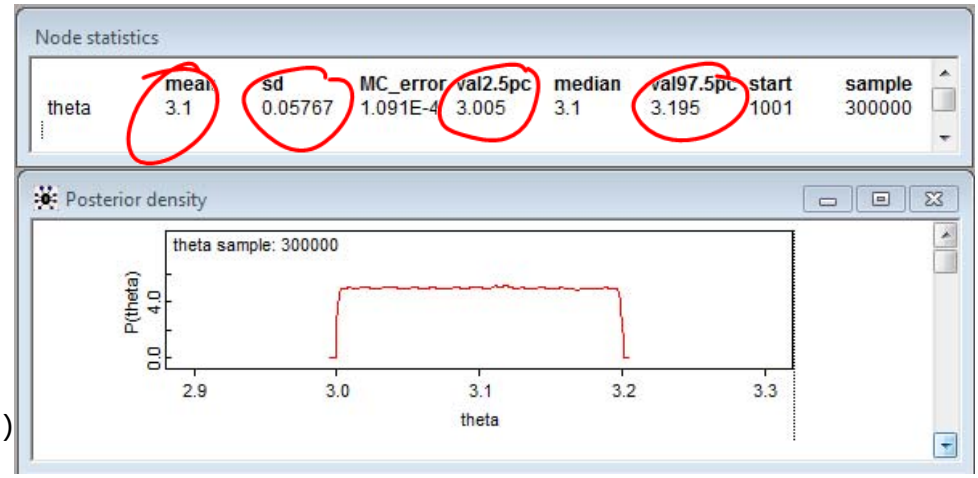
10 pts **4.** Here is some BUGS code and output for a Bayes analysis in a statistical problem involving a parameter θ.

```
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, $f(x|\theta)$:

$f(x|\theta) = I[\theta - .5 < x < \theta + .5]$

The prior density for θ, $g(\theta)$:

$g(\theta) = \frac{1}{20} I[-10 < \theta < 10]$

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

$E[\theta|data] \approx 3.1$

An approximate 95% credible interval for θ:

(3.005, 3.195) has approximately 95% posterior probability

An approximate value for the posterior standard deviation of θ:

$\sqrt{Var(\theta|data)} \approx .0577$

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

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

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

$T(1) =$ 1

$T(2) =$ 2

$T(3) =$ 3

$T(4) =$ 4

$T(5) =$ 2

the original entries in the columns are in the same ratios

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 $R(\theta)=\mathrm{E}_\theta L(\theta, a^{\text{opt}}(x))$.

$a^{\text{opt}}(1) =$ 3

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

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

$a^{\text{opt}}(4) =$ 3

$a^{\text{opt}}(5) =$ 2

by picking the biggest product in each column

$R(1) =$ 1.0

$R(2) =$ .2+.1+.1 = .4

$R(3) =$ .2+.1 = .3

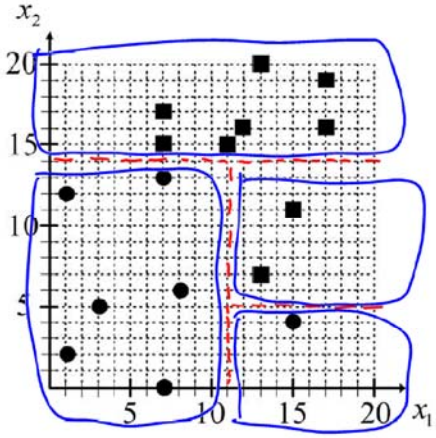
(for what it's worth the Bayes risk is .2(1.0)+.3(.4)+.5(.3) i.e. .47)

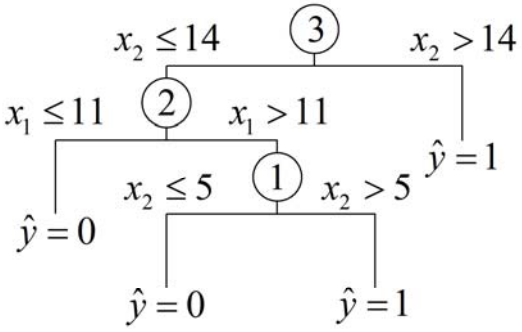
10 pts

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 $\alpha = 4$.





smallest C_4

Tree Pruned at | $\overline{\text{err}}$

None	0 $\quad C_4 = 4 + 4(0)$
(1)	$\frac{1}{16}(1)$ $\quad C_4 = 3 + 4\left(\frac{1}{16}\right) = 3\frac{1}{4}$
(2)	$\frac{1}{16}(2)$ $\quad C_4 = 2 + 4\left(\frac{2}{16}\right) = 2\frac{1}{2}$
(3)	$\frac{1}{16}(7)$ $\quad C_4 = 1 + 4\left(\frac{7}{16}\right) = 2\frac{3}{4}$

The best tree/subtree for $\alpha = 4$ is the one pruned at: node (2)

10 pts

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

T In general terms, larger data sets support the effective use of more complex predictors than do smaller data sets.

F The number of bootstrap samples employed in a bagging predictor (B) is a complexity parameter for bagging. it is, rather, a convergence parameter

F Increasing a bandwidth parameter (λ) in a smoothing algorithm ~~increases~~ decreases the complexity of the corresponding predictor.

F Nearest neighbor and local regression smoothing predictors are particularly ~~effective~~ ineffective for high-dimensional (of the input vectors) prediction problems.

F If one of a set of SEL predictors is essentially $\mathrm{E}[y \mid \mathbf{x}]$, substantial improvement in that predictor can often be obtained through "ensemble" methods using the whole set.

One cannot improve on $E[y|x]$ as a SEL predictor.

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.)

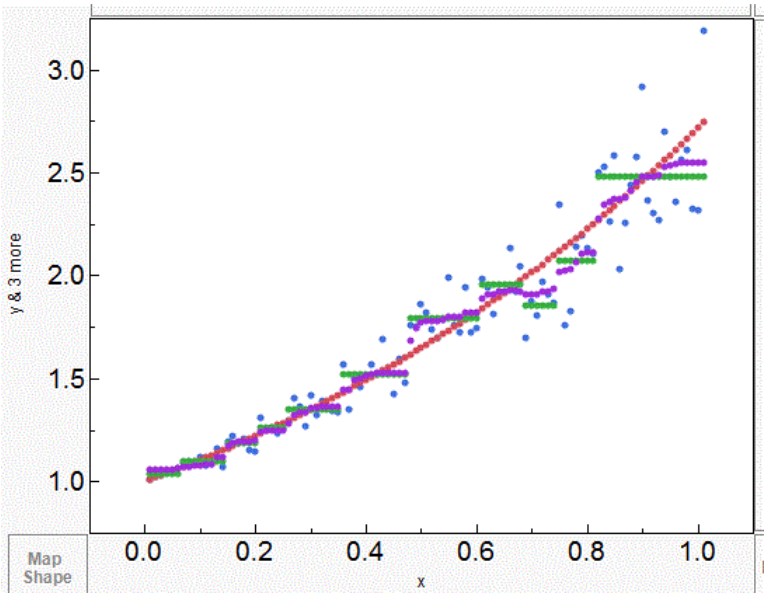
Optimize $Q(\beta) = (-4-\beta(-1))^2 + (-4.5-\beta(-.7))^2 + (2.5-\beta(.7))^2 + (6.0-\beta(1.0))^2 + \beta^2$

$Q'(\beta) = 2(-4+\beta) + 2(-4.5+.7\beta) - 2(2.5-.7\beta) - 2(6-\beta) + 2\beta$

$Q'(\beta) = 0 \Rightarrow (-4-4.5-2.5-12) + (1+.7+.7+1+1)\beta = 0$

$4.4\beta = 23$ i.e. $\hat{\beta} = \frac{23}{4.4} = 5.23$

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?

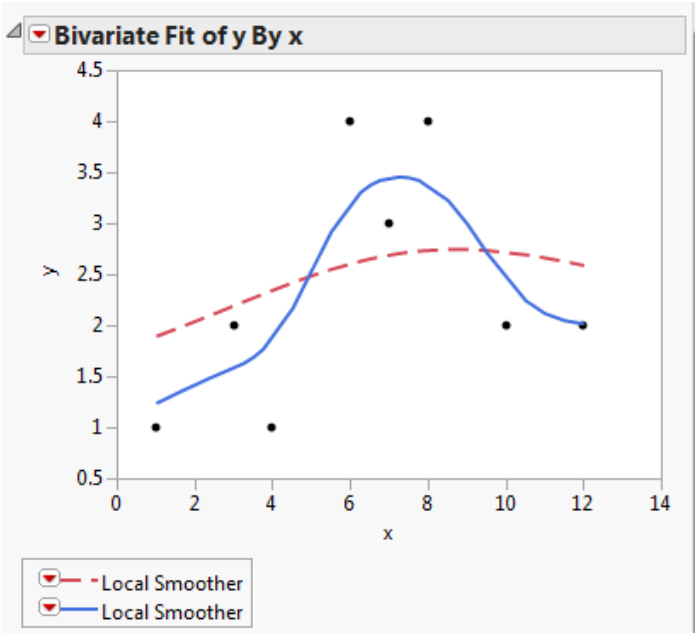
Is this likely just "random/the luck of the draw" for this training set?

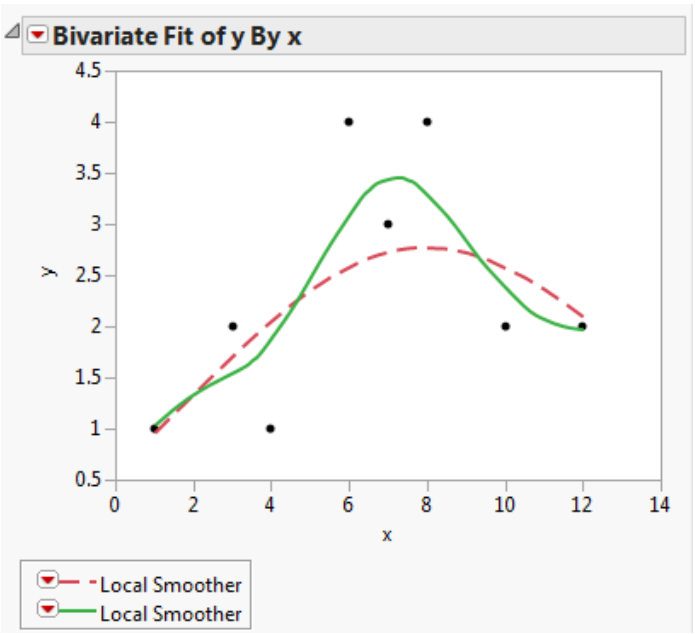
YES or NO (**circle one**)

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.)





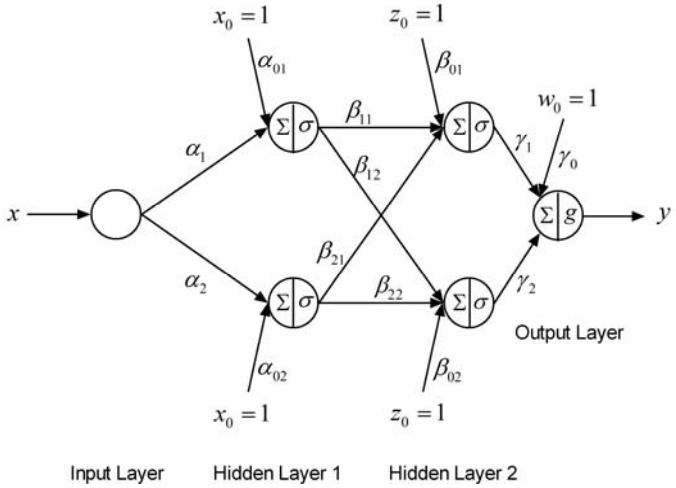
	Degree	Bandwidth
Dashed:	constant/linear	small/large
Solid:	constant /linear	small/large

	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.



$\hat{\alpha}_{01} = 5$ $\quad \hat{\beta}_{11} = 2$ $\quad \hat{\gamma}_1 = -505$

$\hat{\alpha}_1 = -2$ $\quad \hat{\beta}_{12} = -2$ $\quad \hat{\gamma}_2 = 850$

$\hat{\alpha}_{02} = 7$ $\quad \hat{\beta}_{21} = -5$ $\quad \hat{\gamma}_0 = 141$

$\hat{\alpha}_2 = -3$ $\quad \hat{\beta}_{22} = 6$

$\hat{\beta}_{01} = 1$

$\hat{\beta}_{02} = -2$

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 $y = 1$.)