Midterm exam supplement. Closed book and calculators not allowed. Answers may include e^2, $\sqrt{\ }$, etc. but simplfy when possible.

Problem 1.
Consider four classifiers simple classifiers:
Logistic Regression (LR), k Nearest Neighbors (kNN), Naïve Bayes (NB),
and Decision Tree (DT).

Which classifers fit the following statements?

(example of how to answer)
1- Is a classifier. LR, NN, NB, DT

Solution: 1a Recursively partitions the data. DT
Decision tree recursively partitions the data.

1b Has linear decision boundaries. LR
Logistic Regression has linear decision boundaries.

1c Is relatively robust to "the curse of dimensionality". NB,LR,DT
Naïve Bayes and Logistic Regression use strong assumptions (of conditional independence and linearity respectively) to lower their complexity and making them somewhat less prone to overfitting. Typical Decision Tree induction algorithms include feature selection and control of model complexity (stopping criterion used when extending the tree).

1d If given infinite data will converge to optimal classifier. kNN, DT
In principle, k-Nearest Neighbor and Decision Tree are both flexible enough to utilize infinite data to approximate any joint distribution.
kNN's main weakness is probability estimation in a sparse feature space, but infinite data would solve that. For $k = 1$, kNN is still not optimal, but with infinite data approaches optimality with increasing k.
The main weakness of decision tree induction is reduction in effective data size due to repeated partitioning at lower levels of the tree, but infinite data would solve that, in principle allowing a huge decision tree to be induced to approximate any joint distribution of features and class.

Problem 2.
Let $\mathrm{Y}(x)$ be a mixture model of two normal distributions $\mathrm{N}_1(\mu_1, \sigma_1^2)$, and $\mathrm{N}_2(\mu_2, \sigma_2^2)$. With mixture model coefficients (weights) of w_1 and w_2 respectively
$(w_1 + w_2 = 1)$.

Solution:
2a Write down the probability density function for $\mathrm{Y}(x)$.

$$\mathrm{Y}(x) = \frac{w_1\, exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)}{\sqrt{2\pi}\,\sigma_1} + \frac{w_2\, exp\left(\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right)}{\sqrt{2\pi}\,\sigma_2}$$

2b What is the derivative of the natural logarithm of Y(x) with respect to μ_1?

Recall the general relationships:

$$\frac{\partial}{\partial\mu_1}\, ln(\mathrm{Y}(\mu_1, \mu_2, ...)) = \frac{\frac{\partial}{\partial\mu_1}\mathrm{Y}(\mu_1, \mu_2, ...)}{\mathrm{Y}(\mu_1, \mu_2, ...)}$$

We want to compute $\frac{\partial}{\partial\mu_1}\mathrm{Y}(\mu_1, \mu_2, ...)$, which will depend only on the term:

$$\begin{aligned}
\frac{\partial}{\partial\mu_1}\mathrm{Y}(\mu_1, \mu_2, ...) &= w_1 \frac{1}{\sqrt{2\pi}\,\sigma_1} \frac{\partial}{\partial\mu_1}\, exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \\
&= w_1 \frac{1}{\sqrt{2\pi}\,\sigma_1} \frac{2(x-\mu_1)}{2\sigma_1^2}\, exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \\
&= w_1 \frac{x-\mu_1}{\sqrt{2\pi}\,\sigma_1^3}\, exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)
\end{aligned}$$

Combining,

$$\frac{\partial}{\partial \mu_1} ln(\mathrm{Y}(\mu_1, \mu_2, ...)) = \frac{w_1 \frac{x-\mu_1}{\sqrt{2\pi}\,\sigma_1^3}\, exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)}{\frac{w_1\, exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)}{\sqrt{2\pi}\,\sigma_1} + \frac{w_2\, exp\left(\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right)}{\sqrt{2\pi}\,\sigma_2}}$$

$$= \frac{w_1 \frac{z_1}{\sigma_1^2}\, exp\left(\frac{-(z_1^2)}{2}\right)}{\frac{w_1\, exp\left(\frac{-(z_1^2)}{2}\right)}{\sigma_1} + \frac{w_2\, exp\left(\frac{-(z_2^2)}{2}\right)}{\sigma_2}} \qquad z_i \stackrel{\text{def}}{=} \frac{x-\mu_i}{\sigma_i}$$

$$= \frac{w_1 \frac{z_1}{\sigma_1 \sigma_1}\, exp\left(\frac{-(z_1^2)}{2}\right)}{\frac{w_1\, exp\left(\frac{-(z_1^2)}{2}\right)}{\sigma_1} + \frac{w_2\, exp\left(\frac{-(z_2^2)}{2}\right)}{\sigma_2}} \cdot \frac{\frac{\sigma_1}{w_1}}{\frac{\sigma_1}{w_1}}$$

$$= \frac{\frac{z_1}{\sigma_1} exp\left(\frac{-(z_1^2)}{2}\right)}{exp\left(\frac{-(z_1^2)}{2}\right) + \frac{w_2 \sigma_1\, exp\left(\frac{-(z_2^2)}{2}\right)}{w_1 \sigma_2}} \cdot \frac{exp\left(\frac{z_1^2}{2}\right)}{exp\left(\frac{z_1^2}{2}\right)}$$

$$= \frac{\frac{z_1}{\sigma_1}}{1 + \frac{w_2 \sigma_1}{w_1 \sigma_2}\, exp\left(\frac{z_1^2}{2} - \frac{z_2^2}{2}\right)}$$

$$= \frac{\frac{z_1}{\sigma_1}}{1 + \frac{w_2}{w_1} \frac{p(x|\mathrm{N}_2)}{p(x|\mathrm{N}_1)}} \qquad p(x|\mathrm{N}_i) \propto \frac{1}{\sigma_i\, exp(z_i^2/2)}$$

Problem 3.
Consider a random variable X following a Bernoulli distribution over {a,b} with an unknown probability of a, which we denote as $\mathrm{P_a}$. We assume a uniform distribution over $\mathrm{P_a}$. In other words: $\mathrm{p[P_a] = 1,\ \ P_a \in [0,1]}$. Sampling from X we observe the **S**= baa. For convenience we use $\mathrm{F}, \mathrm{F}_a, \mathrm{F}_b$ to denote the length of this sequence and the number of a's and b's it contains.
So for **S**: $\mathrm{F} = 3, \mathrm{F}_a = 2, \mathrm{F}_b = 1$.

For reference, Beta integral:

$$\int_0^1 \mathrm{dP_a\, P_a}^{\mathrm{F}_a}(1-\mathrm{P_a})^{\mathrm{F}_b} = \frac{\Gamma(\mathrm{F}_a+1)\Gamma(\mathrm{F}_b+1)}{\Gamma(\mathrm{F}_a+\mathrm{F}_b+2)} \qquad = \frac{\mathrm{F}_a!\mathrm{F}_b!}{\mathrm{F}_a+\mathrm{F}_b+1!}, \text{for non-negative integers } \mathrm{F}_a, \mathrm{F}_b$$

Solution:
3a. What is likelihood of $\mathrm{P_a}$ given **S**, i.e. $\mathrm{P}[\mathbf{S}|\mathrm{P_a}]$?

$$\mathrm{P}[\mathbf{S}|\mathrm{P_a}] = \mathrm{P_a}^{\mathrm{F}_a}(1-\mathrm{P_a})^{\mathrm{F}_b} = \mathrm{P_a}^2(1-\mathrm{P_a})$$

3b. Recalling that we are using a uniform prior for $\mathrm{P_a}$, what is the posterior probability distribution of $\mathrm{P_a}$ after seeing **S**?

In general `posterior` = `prior` $\times$`likelihood`.
Due to uniform priors, the posterior is proportional to the likelihood, so all that remains is to find a valid probability density proportional to this likelihoo, which turns out to be a beta distribution:

$$\text{beta dist}(\mathrm{F}_a+1, \mathrm{F}_b+1) = \frac{\Gamma(\mathrm{F}_a+1\ +\ \mathrm{F}_b+1)}{\Gamma(\mathrm{F}_a+1)\Gamma(\mathrm{F}_b+1)}\mathrm{P_a}^{\mathrm{F}_a}(1-\mathrm{P_a})^{\mathrm{F}_b} \propto \mathrm{P_a}^{\mathrm{F}_a}(1-\mathrm{P_a})^{\mathrm{F}_b}$$

So the posterior is beta dist$(\mathrm{F}_a+1, \mathrm{F}_b+1)$. In our specific case substituting $\mathrm{F}_a = 2, \mathrm{F}_b = 1$

$$\text{normalization factor} = \frac{\Gamma(\mathrm{F}_a+1\ +\ \mathrm{F}_b+1)}{\Gamma(\mathrm{F}_a+1)\Gamma(\mathrm{F}_b+1)} = \frac{\Gamma(5)}{\Gamma(3)\Gamma(2)} = \frac{4!}{2!1!} = \frac{24}{2} = 12$$

So,

$$\text{probability density } \ p(\mathrm{P_a}|\mathbf{S}) = \mathbf{12}\,\mathrm{P_a}^{\mathbf{2}}(\mathbf{1}-\mathrm{P_a})$$

3c. Given we have seen **S**, what is the probability that the next letter will be a?

Using **Sa** to denote the **S** followed by an a,

$$\begin{aligned}
\int_0^1 p(\mathbf{Sa}|\mathbf{S},\mathrm{P_a})\, p(\mathrm{P_a}|\mathbf{S})\, \mathrm{dP_a} &= \int_0^1 \mathrm{P_a}\, \frac{p(\mathrm{P_a})\, p(\mathbf{S}|\mathrm{P_a})}{p(\mathbf{S})}\, \mathrm{dP_a} && p(\mathbf{Sa}|\mathbf{S},\mathrm{P_a}) = \mathrm{P_a};\ \text{Bayes Law on } p(\mathrm{P_a}|\mathbf{S}) \\
&= \frac{\int_0^1 \mathrm{P_a}\, p(\mathbf{S}|\mathrm{P_a})\, \mathrm{dP_a}}{p(\mathbf{S})} && \text{by uniform prior, } p(\mathrm{P_a}) = 1;\ p(\mathbf{S}) \text{ independent of } \mathrm{P_a} \\
&= \frac{\int_0^1 \mathrm{P_a}\, \mathrm{P_a}^2(1-\mathrm{P_a})\, \mathrm{dP_a}}{\int_0^1 \mathrm{P}(\mathbf{S}|\mathrm{P_a})\mathrm{dP_a}} = \frac{\int_0^1 \mathrm{P_a}^3(1-\mathrm{P_a})\, \mathrm{dP_a}}{\int_0^1 \mathrm{P_a}^2(1-\mathrm{P_a})\, \mathrm{dP_a}} \\
&= \frac{3!1!/5!}{2!1!/4!} = \frac{3!}{2!}\cdot\frac{4!}{5!} = \frac{3}{5} = 0.6
\end{aligned}$$

Problem 4.
Let Y $\sim$ U$(0,1)$ denote the uniformly distribution over [0,1], with probability density function:

$$\mathrm{P}[y = x] = dx; \;\; 0 \leq x \leq 1.$$

Further let M_k denote the random variable obtained by taking the minimum of k independent samples from Y.
So M_2 is the minimum of two samples, etc.

4a. What is the probability density function of M_2?

4b. More generally, what is the probability density function of M_k?

Solution:

This problem nicely illustrates how using Bayes law can help one solve many problems. If P[A|B] stumps you, try working out P[B|A].

First, I assert that the probability density M_k is independent of which sample happened to be the smallest. In other words, $\mathrm{M}_k \stackrel{\text{def}}{=} min(y_1 \ldots y_k)$ is independent of $arg\,min(y_1 \ldots y_k)$.
This seems intuitive to me by symmetry, but we can use Bayes Law to make it more so.
Letting y_m denote the smallest sample;

$$\begin{aligned} p(y_m = x | m = i) \;&=\; \frac{p(y_m = x)\,\mathrm{P}[m = i | y_m = x]}{\mathrm{P}(m = i)} \\ &=\; p[y_m = x]\,\frac{\mathrm{P}[m = i | y_m = x]}{\mathrm{P}[m = i]} \;=\; p[y_m = x]\,\frac{1/k}{1/k} \;=\; p[y_m = x] \end{aligned}$$

Where lower case p denotes probability density, and $\mathrm{P}[m = i | y_m = x]$ is the probability that the smallest sample was i, given that the smallest sample had a value of x.
This demonstrates that knowing which sample is the smallest does not affect the probability distribution of the value of the smallest. For simplicity let's assume the first sample is the smallest, so we now want to figure out $p[y_1 = x | m = 1]$.
Again, using Bayes law we can transform this into an easier question.

$$p[y_1 = x | m = 1] \;=\; \frac{\mathrm{P}[m = 1 | y_1 = x]\,p[y_1 = x]}{\mathrm{P}[m = 1]} \;=\; \frac{\mathrm{P}[m = 1 | y_1 = x]}{1/k}$$

Where $p[y_1 = x] \;=\; 1$, because Y $\sim$ U$(0,1)$.
$m = 1$ when $\forall_{k>1}\; y_k \geq y_1$ — equivalent to $k-1$ flips, all coming up tails, of a coin with probability of heads y_1.
So $\mathrm{P}[m = 1 | y_1 = x] \;=\; (1 - y_1)^{k-1}$.
Combining,

$$p[y_1 = x | m = 1] \;=\; \frac{\mathrm{P}[m = 1 | y_1 = x]}{1/k} \;=\; k(1 - y_1)^{k-1}$$

Problem 5.
An basket contains k balls, of which b are black. One ball is drawn from the basket and then replaced. This is done n times.

Let n_b denote the number of times the ball drawn is black.

Question:

In terms of the variables k, b and n;

Solution: 5a. What is the probability distribution of n_b?

Let p denote the probability of drawing a black ball: $\frac{b}{k}$.

$$n_b \sim \text{binomial dist.(n, p)} = \binom{n}{n_b} p^{n_b}(1-p)^{n-n_b}$$

5b. What is the mean, variance, and standard deviation of this probability distribution?

mean: $\mu = np = n\frac{b}{k}$ variance: $\sigma^2 = n\mathrm{E}[(\mathrm{B}-p)^2]$, where B is an indicator variable defined for a single draw as: B = 1 if the drawn ball is black and, B = 0 otherwise.

$$\begin{aligned}\mathrm{E}[(\mathrm{B}-p)^2] = \mathrm{P}[\mathrm{B}=0](0-p)^2 + \mathrm{P}[\mathrm{B}=1](1-p)^2 &= (1-p)\,(0-p)^2 + p\,(1-p)^2 &&= (1-p)p^2 + p(1-p)^2 \\ &= p(1-p)(p+(1-p)) &&= p(1-p)\end{aligned}$$

So variance is $\sigma^2 = np(1-p) = \mu(1-p)$ and standard deviation is $\sigma = \sqrt{\sigma^2}$.

5c. What is the mean, variance, and standard deviation of n_b for the specific cases of: n=5, b=2, k=10; and n=225, b=2, k=10?

n	p	μ	σ^2	σ
5	0.20	$5 \cdot 0.2 = 1$	$1 \cdot 0.8 = 0.8$	$\sqrt{0.8} \approx 0.894$
225	0.20	$225 \cdot 0.2 = 45$	$45 \cdot 0.8 = 36$	$\sqrt{36} = 6$

Problem 6.
There are five baskets, $u_0 \dots u_4$, each contains 4 balls. Of the four balls they contain, u_0 has no black balls, u_1 has one black ball, ..., finally u_4 has only black balls (4/4).
First, a basket u is picked at random (with $u_0 \dots u_4$ all having an equal chance).

Then 3 balls are randomly drawn from basket u, one at a time, replacing the ball each time.
In other words 3 balls are sampled from u with replacement.
Let b denote the number of the 3 drawn balls which are black.

Question:
In the case where $b = 2$, what is the probability that the next ball drawn will be black?
Show your work.

Solution: First we observe that since both a black and a non-black ball have been drawn, we can eliminate u_0 and u_4 from consideration.

Next we note that the prior probability of picking each basket is uniform, so the posterior probability of the basket will be proportional to the likelihood.
The problem statement gives only the counts (two black and one non-black). I assert that assuming a particular sequence will not change the answer and slightly simplifies the likelihood. So let's assume our data is a particular sequence with two black balls, say (black, black, non-black). Given this data, the likelihood of a basket with probability p is $p^2(1-p)$.
Computing this for the three baskets yields:

	$\mathrm{P}[\text{black}\vert u_i]$	$\mathrm{P}[\text{data}\vert u_i]$	$64\,\mathrm{P}[\text{data}\vert u_i] = 20\,\mathrm{P}[u_i\vert\text{data}]$	$20\,\mathrm{P}[u_i\vert\text{data}]\,4\,\mathrm{P}[\text{black}\vert u_i]$
basket u_i	p	$p^2(1-p)$	$p^2(1-p)/4^3$	
u_1	$\frac{1}{4}$	$\frac{1}{4}^2\,\frac{3}{4}$	$1^2 \cdot 3 = 3$	$3 \cdot 1 = 3$
u_2	$\frac{2}{4}$	$\frac{2}{4}^2\,\frac{2}{4}$	$2^2 \cdot 2 = 8$	$8 \cdot 2 = 16$
u_3	$\frac{3}{4}$	$\frac{3}{4}^2\,\frac{1}{4}$	$3^2 \cdot 1 = 9$	$9 \cdot 3 = 27$
relevant sums		$\frac{20}{64}$	20	$80\,\mathrm{P}[\text{black}\vert\text{data}] = \mathbf{46}$

Answer: $\frac{46}{80} = \frac{23}{40} = 0.575$
Check our work.

$$64\,\mathrm{P}[\text{data}|u_i] = 20\,\mathrm{P}[u_i|\text{data}] \implies \frac{20}{64} = \frac{\mathrm{P}[\text{data}|u_i]}{\mathrm{P}[u_i|\text{data}]} \equiv \frac{\mathrm{P}[\text{data}]}{\mathrm{P}[u_i]}$$

From the problem statement we know $\mathrm{P}[u_i] = \frac{1}{5}$,
so the above implies that $5\mathrm{P}[\text{data}] = \frac{20}{64} \implies \mathrm{P}[\text{data}] = \frac{4}{64} = \frac{1}{16}$.

Let's check. From the table above, we know $\sum_{u_i} \mathrm{P}[\text{data}|u_i] = \frac{20}{64}$.

The joint probability $\mathrm{P}[\text{data}, u_i] \equiv \mathrm{P}[u_i]\,\mathrm{P}[\text{data}|u_i]$,
All $\mathrm{P}[u_i]$ terms are equal so,

$$\mathrm{P}[\text{data}] = \sum_{u_i} \mathrm{P}[u_i]\,\mathrm{P}[\text{data}|u_i] = \sum_{u_i} \frac{1}{5}\mathrm{P}[\text{data}|u_i] = \frac{1}{5}\sum_{u_i} \mathrm{P}[\text{data}|u_i] = \frac{1}{5}\frac{20}{64} = \frac{4}{64} = \frac{1}{16}\checkmark$$