1.
a) Given rule:

var(X|Y) = E[(X − E[X|Y])^2|Y]

Prove var(X) = E[var(X|Y)] + var(E[X|Y])

Starting with RHS, we have:
E[var(X|Y)] + var(E[X|Y])

(i)
$E[var(X|Y)] = E\big[E[(X - E[X|Y])^2|Y]\big]$ // Using given rule
$= E[E[X^2|Y] - E[X|Y]^2] = E\big[E[X^2|Y]\big] = E[E[X|Y]^2]$ // Linearity of Expectations
$= E[X^2] - E[E[X|Y]^2]$ // Law of Iterated Expectations

(ii)
$var(E[X|Y]) = E[E[X|Y]^2] - E\big[E[X|Y]\big]^2$ // Definition of variance
$= E[E[X|Y]^2] - E[X]^2$ // Law of Iterated Expectations

Add parts (i) and (ii) together.
$E[X^2] - E[X]^2 = var(X)$ // Definition of variance

b) Recall the proven formula in part a: var(X) = E[var(X|Y)] + var(E[X|Y])

Here, $X = \Sigma_{i=1}^{N} Z_i$ and $Y = N$.

(i)
$var(X|Y) = var(\Sigma_{i=1}^{N} Z_i|N) = var(\Sigma_{i=1}^{N} Z_i\big|N) = var(\Sigma_{i=1}^{N} Z_i\big|N)$ // Condition on N
$= var(Z_1 + Z_2 + \cdots + Z_N) = \Sigma_{i=1}^{N} var(Z_i)$ // Variance of sum of iid RV
$\Rightarrow E[var(X|Y)] = E[\Sigma_{i=1}^{N} var(Z_i)\ ] = E[N * var(Z)]$ // Placing into expected value and using fact that all Z_i come from same distribution and has same variance
$= E[N]E[var(Z)] = E[N]var(Z)$ // Simplify

(ii)
$E[X|N] = E[\Sigma_{i=1}^{N} Z_i|N] = \Sigma_{i=1}^{N} E[Z_i|N] = NE[Z]$ // Linearity of Expectations
$\Rightarrow var(E[X|Y]) = var(NE[Z]) = E[Z]^2 var(N)$ // Variance of Multiplied Number

Add parts (i) and (ii) together.
$var(X) = var(\Sigma_{i=1}^{N} Z_i)$ = E[N]var(Z) + (E[Z])^{2}var(N)

2. As a base case, suppose there is only $n = 1$ coins. Since there is only one coin to flip, that very coin must be the one fair coin with $p = q = \frac{1}{2}$, where p and q denote the probability for a specific

coin to turn heads or tails, respectively. The number of even heads can only be zero i.e. the coin turns tail. That means $P_1 = p = \frac{1}{2}$, where P_1 denotes the probability of getting an even number of heads in a total of 1 coin flip.

Now suppose there is a total of an arbitrary *n* coins. It follows that:
$P_n = P_{n-1} * q + (1 - P_{n-1}) * p = P_{n-1} * (1 - p) + (1 - P_{n-1}) * p = P_{n-1} + p * (1 - 2P_{n-1})$

This says that the probability of even coins being heads in *n* flips is the probability of *n*-1 coins having even heads with the n^{th} coin being tails or *n*-1 coins having odd heads with the n^{th} coin being heads to make a total of even number of heads.

Suppose that the fair coin $\left(p = q = \frac{1}{2}\right)$ is already present in the first *n*-1 coins. This would indicate $P_{n-1} = \frac{1}{2} \Rightarrow P_n = \frac{1}{2} + p * \left(1 - 2 * \frac{1}{2}\right) = \frac{1}{2}$. Recall for $n = 2$, the corresponding $P_{n-1} = P_1 = \frac{1}{2}$ as proven before in the base case. By induction, $P_{n-1} = \frac{1}{2}$ for $n = 3,4,$ ... etc. The other scenario is if the fair coin is the last, or n^{th} , coin. This means that P_{n-1} is not necessarily, and most likely not, $\frac{1}{2}$. The probability for even number of heads now is $P_n = P_{n-1} + \frac{1}{2} * (1 - 2P_{n-1}) = P_{n-1} + \frac{1}{2} - P_{n-1} = \frac{1}{2}$. Therefore, regardless of the position that the fair coin comes in the sequence of *n* coins, the probability of the even heads is still $\frac{1}{2}$.

3.
a) $P(X > Y) \Rightarrow P(X - Y > 0)$
$\Rightarrow \int_{y=0}^{\infty} \int_{x=y}^{\infty} f(x)f(y)dxdy = \int_{y=0}^{\infty} \int_{x=y}^{\infty} \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 y} dxdy$ // Joint probability
$= \lambda_1 \lambda_2 \int_{y=0}^{\infty} \int_{x=y}^{\infty} e^{-\lambda_1 x - \lambda_2 y} dxdy = \lambda_1 \lambda_2 \int_{y=0}^{\infty} [-\frac{1}{\lambda_1} e^{-\lambda_1 x - \lambda_2 y}|_{x=y}^{\infty}]dy$ // Inner integral
$= -\lambda_2 \int_{y=0}^{\infty} -e^{-\lambda_1 y - \lambda_2 y}\, dy = \lambda_2 \int_{y=0}^{\infty} e^{-(\lambda_1 + \lambda_2)y}\, dy$ // Simplify previous step
$= \frac{-\lambda_2}{\lambda_1 + \lambda_2} \left[e^{-(\lambda_1+\lambda_2)y}|_{y=0}^{\infty}\right] = \frac{-\lambda_2}{\lambda_1+\lambda_2}[0 - 1] = \frac{\lambda_2}{\lambda_1+\lambda_2}$

b) Let *t* denote an arbitrary value such that both $X > t$ and $Y > t$ are true. Then logically, the minimum between *X* and *Y* is also greater than *t*.

$\mathrm{P(Z > t) = P(min\{X, Y\} > t) = \ P(X > t, Y > t) = P(X > t) * P(Y > t) = \ e}^{-\lambda_1 t} * \mathrm{e}^{-\lambda_2 t}$
$= \mathrm{e}^{-(\lambda_1+\lambda_2)t}$

The CDF would be $F_Z(t) = P(Z \leq t) = 1 - P(Z > t) = 1 - e^{-(\lambda_1+\lambda_2)t}$.
Following the form for CDF of the exponential distribution, the mean is easily seen being $\frac{1}{\lambda_1+\lambda_2}$.

c) $P(Z|Z = X) = P(Z|X \leq Y) = \frac{P(Z,X\leq Y)}{P(X\leq Y)}$ // Definition of conditional probability

(i) numerator
$P(Z, X \leq Y) = P(Z, X = \min\{X, Y\}) = (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)z}$ // Based on CDF computed from part (b)

(ii) denominator

$P(X \leq Y) = \int_{y=0}^{\infty} \int_{x=0}^{y} f(x)\,f(y)\,dx = \int_{y=0}^{\infty} \int_{x=0}^{y} \lambda_1 e^{-\lambda_1 x} \lambda_2 e^{-\lambda_2 y}\,dxdy =$
$\lambda_1\lambda_2 \int_{y=0}^{\infty} \int_{x=0}^{y} e^{-\lambda_1 x - \lambda_2 y}\,dxdy$
$= -\lambda_2 \int_{y=0}^{\infty} e^{-\lambda_1 x - \lambda_2 y}\,|_{x=0}^{y} dy = -\lambda_2 \int_{y=0}^{\infty} e^{-(\lambda_1+\lambda_2)y} - e^{-\lambda_2 y} dy$
$= -\lambda_2 \left[\left(-\frac{1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)y} + \frac{1}{\lambda_2} e^{-\lambda_2 y}\right)|_{y=0}^{\infty}\right] = -\lambda_2 \left[\frac{1}{\lambda_1+\lambda_2} - \frac{1}{\lambda_2}\right] = 1 - \frac{\lambda_2}{\lambda_1+\lambda_2} = \frac{\lambda_1}{\lambda_1+\lambda_2}$
Note that in part (a), we proved that $P(X > Y) = \frac{\lambda_2}{\lambda_1+\lambda_2}$. The denominator can be verified by $P(X \leq Y) = 1 - P(X > Y)$.

Combine numerator and denominator into fraction.
$P(Z|Z = X) = \frac{(\lambda_1+\lambda_2)e^{-(\lambda_1+\lambda_2)z}}{\frac{\lambda_1}{\lambda_1+\lambda_2}} = \frac{(\lambda_1+\lambda_2)^2}{\lambda_1} e^{-(\lambda_1+\lambda_2)z}$
Since it is conditioned that $Z = X$, the above equation can also replace *z* with *x*.

d) Just like in part (c), conditioning on $Z = X$ would only consider a distribution when $X \leq Y$. Similar to part (c), a convolution of the individual density functions, in this case denoted by $f(y)f(z)$ as the integrand, would be computer such that $X \leq Y$ is taken into account. This would yield another exponential distribution in the end by looking at the previous part.

Due to the memoryless property of exponential distribution, the difference $Y - Z$ comes from the same distribution as *Y*, without needing to known where the minimum i.e. *X* even occurred prior to *Y*. Since distribution of *Y* is known to have mean $\frac{1}{\lambda_2}$, this is also the case for $Y - Z$.

4.
a) $E[X|X < c] = \frac{1}{P(X<c)} \int_0^c x f(x) I(x < c) dx = \frac{1}{P(X<c)} \int_0^c x\lambda e^{-\lambda x}\,dx = \frac{1}{P(X<c)} \lambda \int_0^c x e^{-\lambda x}\,dx$
//Definition of EV
$= \frac{1}{P(X<c)} \lambda \left[-\frac{x}{\lambda} e^{-\lambda x}|_0^c - \int_0^c -\frac{1}{\lambda} e^{-\lambda u}\,du\right] = \frac{1}{P(X<c)} \lambda \left[-\frac{c}{\lambda} e^{-\lambda c} + \frac{1}{\lambda} \int_0^c e^{-\lambda u}\,du\right]$ // Integration by parts
$= \frac{1}{P(X<c)} \lambda \left[-\frac{c}{\lambda} e^{-\lambda c} - \frac{1}{\lambda^2} e^{-\lambda u}|_0^c\right] \qquad = \frac{1}{P(X<c)} \left(-ce^{-\lambda c} - \frac{1}{\lambda}\left(e^{-\lambda c} - 1\right)\right) = \frac{1}{P(X<c)} \left(\frac{1}{\lambda} - \frac{1}{\lambda} e^{-\lambda c} - ce^{-\lambda c}\right)$
Note that $P(X < c) = F_X(c) = 1 - e^{-\lambda c}$
$\Rightarrow E[X|X < c] = \frac{\frac{1}{\lambda} - \frac{1}{\lambda} e^{-2c\lambda} - ce^{-2c\lambda}}{1 - e^{-\lambda c}}$

b) Use the identity: $\mathrm{E[X] = E[X|X < c] * P(X < c) + E[X|X > c] * P(X > c)}$

The identity can be re-arranged to compute $E[X|X < c] = \frac{E[X]-E[X|X>c]*P(X>c)}{P(X<c)}$

The following parts are known or computed using properties of the exponential distribution.

$E[X] = \frac{1}{\lambda}$

$E[X|X > c] = \int_c^\infty xf(x)dx = \int_c^\infty x\lambda e^{-\lambda x}\, dx = \lambda \int_c^\infty xe^{-\lambda x}\, dx$ // Definition of expected value

$= \lambda \left[-\frac{x}{\lambda} e^{-\lambda x}|_c^\infty - \int_c^\infty -\frac{1}{\lambda} e^{-\lambda u}\, du\right] = \lambda \left[\frac{c}{\lambda} e^{-\lambda c} + \frac{1}{\lambda}\int_c^\infty e^{-\lambda u}\right] = \lambda \left[\frac{c}{\lambda} e^{-\lambda c} - \frac{1}{\lambda^2} e^{-\lambda u}|_c^\infty\right]$ // Integration by parts

$= \lambda \left[\frac{c}{\lambda} e^{-\lambda c} + \frac{1}{\lambda^2} e^{-\lambda c}\right] = ce^{-\lambda c} + \frac{1}{\lambda} e^{-\lambda c}$

$P(X > c) = 1 - P(x \leq c) = 1 - F_X(c) = 1 - \left(1 - e^{-\lambda c}\right) = e^{-\lambda c}$

$P(X < c) = F_X(c) = 1 - e^{-\lambda c}$

Putting the formula together.

$E[X|X < c] = \frac{\frac{1}{\lambda} - \left(ce^{-\lambda c} + \frac{1}{\lambda} e^{-\lambda c}\right)\left(e^{-\lambda c}\right)}{1-e^{-\lambda c}} = \frac{\frac{1}{\lambda} - \frac{1}{\lambda} e^{-2c\lambda} - ce^{-2c\lambda}}{1-e^{-\lambda c}}$

5.

a) Because of the memoryless property of the exponential distribution, the first person between A and B to be done being served still has mean $\frac{1}{\lambda}$ at the point C begins being served. Person C also has mean $\frac{1}{\lambda}$ during the start of being served, as stated by the problem. Since both the first person to be finished and person C have the same mean time being server i.e. $\frac{1}{\lambda}$, the probability that person C leaves last is $\frac{1}{2}$.

b) Let R1 and R2 be the remaining time left for clerks 1 and 2 with their customers.

The third customer, C, would leave after addition time spent on top of the first of 2 customers to be finished.

$\mathrm{E[T\ |R2 \ \leq\ R1]} = \frac{1}{\lambda_1+\lambda_2} + \frac{1}{\lambda_2} \Rightarrow \frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{3}{2\lambda}$

c) $P\left(\max\{A, B\} > \min\{A, B\} + \frac{1}{\lambda}\right) = \frac{2}{5\lambda}$

6.

a) Removing the assumption of independent features, the same model becomes:

Let $t = x_i$.

$P(X = t|Y = c, \pi, \theta) = P(X = x_i|Y = c) = P(x_{i1}|Y = c) * P(X = x_{i2}|Y = c, x_{i1}) * \cdots * P(x_{iD}|Y = c, x_{i\,1:D-1}) = \Pi_{j=1}^{D} P(x_{ij}|y = c, x_{i1}, \ldots, x_{i\,j-1})\ for\ c = y_i$

With the assumption of independent features, we can arrive at $P(X|Y = c)$ by simply multiply the feature weights of D features together per class, yielding CD parameters. We cannot do that now since independence is not valid for feature. For each of C classes, we can instead lookup $P(X|Y = c)$ given using the D features as a binary index or binary string as key. For a feature space of length D, there are 2^D unique binary indexes that are possible. Therefore, the number of parameters over C classes is $C * 2^D$.

b) For one instance x_i:

$P(X = x_i, Y = y_i|\pi, \theta) = P(Y = y_i|\pi, \theta) P(X = t|Y = c, \pi, \theta) = \pi_C [\Pi_{j=1}^{D} P(x_{ij}|y = c, x_{i1}, \ldots, x_{i\,j-1})]$
$= \frac{1}{C} \Pi_{j=1}^{D} P(x_{ij}|y = c, x_{i1}, \ldots, x_{i\,j-1})\ for\ c = y_i$

Note that each instance x_i has its own unique θ_i that is looked up in the table described in part (a). Because features are not iid random variables, θ_i cannot be broken down into a product along feature subscript *j* as seen in NBC.

Likelihood function:

$P(\mathcal{D}|\pi, \theta) = \Pi_{i=1}^{N} P(X = x_i, Y = y_i|\pi, \theta) = \Pi_{i=1}^{N} P(Y = y_i|\pi, \theta) P(X = t|Y = c, \pi, \theta) =$
$\Pi_{i=1}^{N} [\frac{1}{C} \Pi_{c=1}^{C} \Pi_{j=1}^{D} P(x_{ij}|y = c, x_{i1}, \ldots, x_{i\,j-1})]$
$= \frac{1}{C^N} \Pi_{i=1}^{N} \Pi_{c=1}^{C} \Pi_{j=1}^{D} P(x_{ij}|y = c, x_{i1}, \ldots, x_{i\,j-1}) = \frac{1}{C^N} \Pi_{i=1}^{N} \theta_i$

There can only be one specific D-bit vector and class per instance x_i, therefore the two (2) inner product operators are dismissed for simplification.

Log-likelihood function:

$\log P(\mathcal{D}|\pi, \theta) = \log\left(\frac{1}{C^N} \Pi_{i=1}^{N} \theta_i\right)$
$= \log\left(\frac{1}{C^N}\right) + \Sigma_{i=1}^{N} \log \theta_i$
$= -N \log C + \Sigma_{i=1}^{N} \log \theta_i$

c) The class distribution is given as uniform, so $\hat{\pi} = \frac{1}{C}$.

$\hat{\theta}_i = \max_{\theta_i \in [0,1]} (-N \log C + \Sigma_{i=1}^{N} \log \theta_i) = \max_{\theta_i \in [0,1]} \log \theta_i = 1$

d) Runtime complexity is $O(ND)$ for both NBC and the full model. Both algorithms require iteration through N instances, under which there is an iteration through D features. Note that this accounts for the formulation of binary index in $O(D)$ time.

Memory space is $O(CDN)$ for NBC. θ_{ij} can be stored in a $D \times C$ matrix. And then there is the dataset of N instances. For the full model, memory space is $O(C * 2^D * N)$. The only difference here is that for each class table, there can be up to 2^D binary indexes.

e) To use plug-in approximation, test each of C classes to find
$\arg\max_{c\in\{1,\ldots,C\}} \mathrm{P}(\mathrm{Y} = \mathrm{c}|\mathrm{X} = \mathrm{x}, \hat{\pi}_{\mathrm{MAP}}, \hat{\theta}_{\mathrm{MAP}})$

We are already given that $\hat{\pi} = \frac{1}{C}$, so no need to compute $\hat{\pi}_{MAP}$ from scratch.
Knowing $\mathrm{P}(\pi, \theta|\mathrm{D}) \propto \mathrm{P}(\mathrm{D}|\pi, \theta)\mathrm{P}(\pi, \theta)$, find estimate $\hat{\theta}_{MAP}$ that maximizes the expression, which can be re-expressed below.

$\mathrm{P}(\pi, \theta|\mathrm{D}) \propto \frac{1}{C^N} \Pi_{i=1}^{N} \theta_i$

Maximizing $\theta_i\ s.t. \theta_i \in [0,1]$ yields 1 as before.

For a single instance to classify, runtime is $O(CD)$ for NBC and $O(\mathrm{C} * 2^{\mathrm{D}})$ for full model.

7.
a) Note that we can use the formula $P(X = i|D) = \frac{a_i+N_i}{\Sigma_{k=1}^{K}(a_k+N_k)}$.
$P(X_{2001} = \mathrm{e}|\mathcal{D}) = \frac{10+260}{270+2000} = \frac{270}{2270} \approx 0.119$

b) It can be shown that $P(X|D) = \frac{B(\alpha+N+X)}{B(\alpha+N)} = \frac{\Pi_{k=1}^{K}\Gamma(a_k+N_k+x_k)\Gamma(\sum k\alpha k+Nk)}{\Gamma(\sum_{k=1}^{K} a_k+N_k+x_k)\prod_{k=1}^{K} \Gamma(\alpha_k+N_k)}$
Using this formula, we have
$P(x_{2001} = p, x_{2002} = a|\mathcal{D}) = P(X|D) = \frac{\Gamma(111)\Gamma(98)\Gamma(2270)}{\Gamma(110)\Gamma(97)\Gamma(2272)} = \frac{(110!)(97!)(2269!)}{(109!)(96!)(2271!)} \approx 0.002$