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

Your Name: ______________________________

Problem 1.
Recall that a Gaussian prior is conjugate to the mean of a Gaussian distribution.
Given:
1. a random variable X is distributed normally given its mean, i.e. $X|\mu \sim (\mu_x, 1)$
2. our prior belief regarding μ is a standard normal: $\mu \sim (0, 1)$
3. we have one data point $x_1 = 10$.

Question: what is the posterior distribution of μ after observing x_1?

1a. Informally justify your answer (可以用中文)
1b. (Challenging?) Mathematically prove your answer.

Solution: This is a mean problem. Seriously, thinking clearly about "the mean of a mean" is part of the challenge here. Let μ_X denote the mean of the data and $u_0 = 0$ denote the mean of our prior estimate of μ_X.

$$p[\mu_x|x_1] = p[\mu_x]\,p[x_1|\mu_x] = \quad \mu_x{\sim}(u_0,1) \;\cdot\; x_1{\sim}(\mu_x,1) \propto \exp\left(\frac{-(\mu_x - u_0)^2}{2}\right)\exp\left(\frac{-(x_1 - \mu_x)^2}{2}\right)$$

Note that by using $\propto$, I have omitted terms independent of μ_x.
This equation is symmetric in terms of μ_0 and x_1,
so my intuition is that the result will have a mean of $\frac{\mu_0+x_1}{2} = \frac{x_1}{2}$.
The question states that a Gaussian prior is conjugate to the mean of a Gaussian distribution, So the trick is to see if the posterior is equivalent to $(\mu_{\texttt{post}}, \sigma^2_{\texttt{post}})$, where $\mu_{\texttt{post}}, \sigma^2_{\texttt{post}}$ are the mean and variation of the posterior estimate of μ_x.

$$p[\mu_x|x_1] = p[\mu_x]\,p[x_1|\mu_x] \propto \quad \exp\left(\frac{-\mu_x^2 - (10 - \mu_x)^2}{2}\right) \propto \quad \exp\left(\frac{-(\mu_{\texttt{post}} - \mu_x)^2}{2\,\sigma^2_{\texttt{post}}}\right)??$$

Repeating the above equation, with substitutions $\mu_0 \to 0, x_1 \to 10$:

$$p[\mu_x|x_1] = p[\mu_x]\,p[x_1|\mu_x] \propto \quad \exp\left(\frac{-\mu_x^2}{2}\right)\exp\left(\frac{-(10 - \mu_x)^2}{2}\right) = \quad \exp\left(\frac{-\mu_x^2 - (10 - \mu_x)^2}{2}\right)$$

Next we try to simplify the exponent:

$$-\mu_x^2 - (10 - \mu_x)^2 = -1\left(\mu_x^2 + (10 - \mu_x)^2\right) = \quad -1(2\mu_x^2 - 2\cdot 10\,\mu_x + 10^2) = \quad -1(2\mu_x^2 - 20\,\mu_x + 100)$$

Let's also consider my guess of $\frac{10}{2} = 5$.

$$-(\mu_x - 5)^2 = -1(\mu_x^2 - 10\,\mu_x + 25) = -½(2\mu_x^2 - 20\,\mu_x + 50)$$

Using the hint from the corresponding terms in blue.

$$\begin{aligned}
\exp\left(\frac{-\mu_x^2 - (10 - \mu_x)^2}{2}\right) &= \exp\left(\frac{-1(2\mu_x^2 - 20\mu_x + 100)}{2}\right) \\
&= \exp\left(\frac{-½(2\mu_x^2 - 20\mu_x + 100)}{½ \cdot 2}\right) \\
&= \exp\left(\frac{-½(2\mu_x^2 - 20\mu_x + 50)}{½ \cdot 2}\right)\exp\left(\frac{-½ \cdot 50}{½ \cdot 2}\right) \\
&= \exp\left(\frac{-(\mu_x - 5)^2}{½ \cdot 2}\right)\exp\left(\frac{-½ \cdot 50}{½ \cdot 2}\right) \propto \exp\left(\frac{-(\mu_x - 5)^2}{½ \cdot 2}\right) \propto \mu_x{\sim}(5, ½)
\end{aligned}$$

So the posterior distribution of μ_x should follow a normal distribution with mean 5 and variance ½.

Your Name: ______________________________

Problem 2.
Imagine rolling a (not necessarily fair) 4-sided die, numbered {1,2,3,4}.
Given:
1. Prior: Your prior belief on the probability of each side is Dirichlet(½, ½, ½, ½).
2. Data: You roll the die twice, getting a 1 and a 3.

Question:
2a. What is the posterior distribution over {1,2,3,4} after observing the data?

Solution:

$$P[\texttt{die}|\texttt{data}] = P[\texttt{die}] \cdot P[\texttt{data}|\texttt{die}] \propto \quad \text{Dirichlet}(½, ½, ½, ½) \cdot p_1\, p_3 \propto \quad \text{Dirichlet}(\frac{3}{2}, \frac{1}{2}, \frac{3}{2}, \frac{1}{2})$$

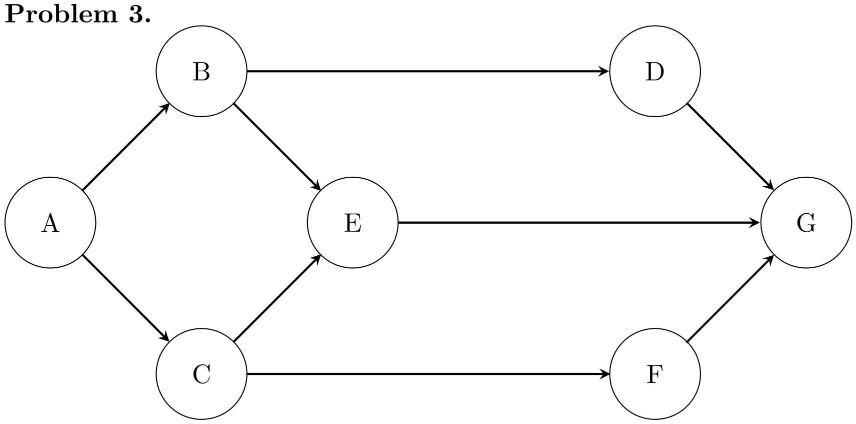
Where "$P[\texttt{die}]$" denotes the prior estimate of the innate probability of {1,2,3,4} of the die; and I have used $\propto$ so that I can omit normalization terms.
2b. What is the probability that the next die roll yields a 3?

Solution: By the "pseudo-counts" method, the probability of a 3 is:

$$P[3] = \frac{1 + ½}{2 + 4 \cdot ½} = \frac{1.5}{4} = \frac{3}{8}$$

Your Name: ____________________

Problem 3.



The above graph is a Bayesian network (aka Belief Network, or probabilistic graphical model). Consider the $\binom{7}{3} = 35$ possible triples of nodes in alphabetical order (A,B,C); (A,B,D); ... ; (E,F,G).

Question:
List the triples (X,Y,Z) for which X and Y are conditionally independent given Z.
Where $X, Y, Z \in \{A, \ldots, G\}, X \neq Y, X \neq Z, Y \neq Z$.

Solution: First let's cut down on the number of (X,Y) pairs. According to the "alphabetical order" condition X and Y cannot be G.
Also note that a direct edge X → Y indicates a dependency between X and Y which cannot be "blocked" by any other node. This immediately precludes the pairs: {AB, AC, BD, BE, CE, CF} being conditionally independent.
For the remaining possibilities we consider more general rules. A Bayesian network guarantees $X \perp Y \,|Z$, the conditional independence of nodes X,Y given node Z
iff all (undirected) paths in from X to Y match one of the following three patterns.

1. X → Z → Y or Y → Z → X
2. X ← Z → Y
3. X → W ← Y where W ≠ Z, nor is W descendent from Z.

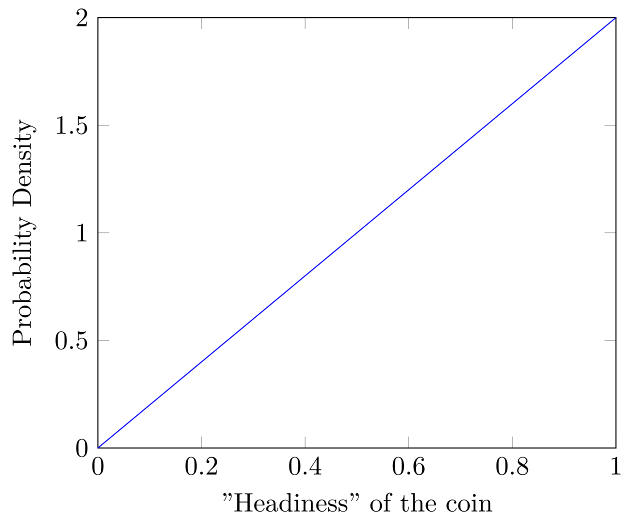
So the game is to find a counter-example path from X to Y which does not match any of the above rules.

XY|Z	Counter Example Path	Reason
AD|E	A → B → D	simple chain without E
AD|F	A → B → D	simple chain without F
AD|G	A → B → D	simple chain without G
AE|F	A → B → E	simple chain without F
AE|G	A → B → E	simple chain without G
AF|G	A → C → F	simple chain without G
BC|D	B ← A → C	B and C have common parent ≠ D
BC|E	B → E ← C	→*← node E is E itself
BC|F	B ← A → C	B and C have common parent ≠ F
BC|G	B → D → G ← F ← C	→*← node G is G itself
BF|G	B → D → G ← F	→*← node G is G itself
CD|E	C → F → G ← D	→*← node is G is a descendant of E
CD|F	C → F → G ← D	→*← node is G is a descendant of F
CD|G	C → F → G ← D	→*← node G is G itself
DE|F	D → G ← E	→*← node G is a descendant of F
DE|G	D → G ← E	→*← node G is G itself
DF|G	D → G ← F	→*← node G is G itself
EF|G	E → G ← F	→*← node G is G itself

Answer: None of the triples fulfill $X \perp Y \,|Z$.

Your Name: ____________________

Problem 4.



This is a coin flipping problem.
Recall a beta distribution is defined as:

$$P_H \sim \mathrm{BetaDist}(a, b) \quad \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, P_H^{a-1}(1-P_H)^{b-1}$$

Given:
1. the data is a single coin toss, yielding "heads".
2. a beta distribution $\mathrm{BetaDist}(a, b)$ was used as a prior.
3. the posterior distribution is as plotted above.

Question:
What were the parameters (a, b) of the beta distribution prior?

Solution: The data is a single head so the likelihood is proportional to P_H.
The posterior distribution plotted in the figure above is a straight line with the probability density proportional to P_H. So, `posterior` $\propto P_H$ and `likelihood` $\propto P_H$.

$$\texttt{prior} = \texttt{posterior}/\texttt{likelihood} = \text{constant}.$$

Therefore the prior must have been $\mathrm{BetaDist}(1, 1)$, since these are the only parameters to BetaDist which give a constant density.

$$\mathrm{BetaDist}(1, 1) \quad \frac{\Gamma(1+1)}{\Gamma(1)\,\Gamma(1)}\, P_H^{1-1}(1-P_H)^{1-1} = \frac{1}{1 \cdot 1}\, P_H^{0}\,(1-P_H)^{0} = 1$$

Your Name: ______________________________

Problem 5.
Dataset:

Class	F_1	F_2	F_3
A	good	good	okay
A	bad	bad	good
A	bad	okay	okay
A	okay	okay	good
A	bad	okay	good
B	good	okay	okay
B	okay	okay	bad
B	okay	good	bad
B	good	bad	bad

Question:
Specify a Naïve Bayes classifer based on the above dataset.
Your classifier should provide enough information to compute the numerical value of $P[class = \mathbf{A} | F_1, F_2, F_3]$ for all 27 combinations of $(F_1, F_2, F_3) \in \{\text{good}, \text{okay}, \text{bad}\}$.
Explicitly state all priors used.

Solution: This is the preferred solution, using pseudo-counts to reflect prior distributions. Let's start by writing the defining equation for Naïve Bayes

$$P[C|F_1, F_2, F_3] \propto P[C]\ P[F_1, F_2, F_3|C] \approx\ P[C]\ P[F_1|C]\ P[F_2|C]\ P[F_3|C]$$

So the parameters needed are $P[C]$, and the three $P[F|C]$ terms.
The problem says to "explicitly state all priors used". I will choose to use a BetaDist$(0.5, 0.5)$ as prior for $P[C]$. The data on $P[C]$ is five **A**s out of total of 9 data items, so using "pseudo-counts" the posterior probability of $P[C]$ would be BetaDist$(0.5 + 5, 0.5 + 4)$ = BetaDist$(5.5, 4.5)$. Each feature has three possible values $\{\texttt{good}, \texttt{okay}, \texttt{bad}\}$, so it is useful to use a Dirichlet distribution as a prior. Here I will assume Dirichlet $0.5, 0.5, 0.5$ for all 3x2=6 possible combinations of feature and class.
It would be in the spirit of this course to derive the posterior distribution of $P[C|F_1, F_2, F_3]$. However I did not go through that in class and I doubt most people do Naïve Bayes classifiers that way.
So instead let's use these priors less ambitiously — as a way to obtain *maximum a posterior* (MAP) estimates.
Using pseudo-counts and the data counts of 5 class **A** out of 9 samples,
we can compute the MAP estimate of $P[C = \mathbf{A}] = \frac{5+0.5}{9+1} = 0.55$.
Likewise the MAP estimate of $P[F|C]$ is given by:

$$\text{MAP estimate } P[F_i|C_j] = \frac{N(F_i, C_j) + 0.5}{N(C_j) + 3 \cdot 0.5} = \frac{2N(F_i, C_j) + 1}{2N(C_j) + 3}$$

Where $N(\texttt{X})$ denotes the count of $\texttt{X}$ in the data. The following table summarizes those counts.

	feature F_1				feature F_2				feature F_3			
	class **A**		class **B**		class **A**		class **B**		class **A**		class **B**	
	good	okay	good	okay	good	okay	good	okay	good	okay	good	okay
$N(F_i\|C_j)$	1	1	2	2	1	3	1	2	3	2	0	1
$P[F_i\|C_j]$	$\frac{3}{13}$	$\frac{3}{13}$	$\frac{5}{11}$	$\frac{5}{11}$	$\frac{3}{13}$	$\frac{7}{13}$	$\frac{3}{11}$	$\frac{5}{11}$	$\frac{7}{13}$	$\frac{5}{13}$	$\frac{1}{11}$	$\frac{3}{11}$

Where $N[\texttt{X}]$ denotes the number of $\texttt{X}$ observed in the data.
And $P[\texttt{bad}|C]$ can be computed as $1 - P[\texttt{good}|C] - P[\texttt{okay}|C]$.
Here I list an example to demonstrate that we have defined enough parameters.

$$\begin{aligned}
P[C = \mathbf{A}|(\texttt{good}, \texttt{bad}, \texttt{okay})] &\propto P[C = \texttt{good}]\ P[F_1 = \texttt{good}|\mathbf{A}]\ P[F_2 = \texttt{bad}|\mathbf{A}]\ P[F_3 = \texttt{okay}|\mathbf{A}] \\
&= 0.55 \cdot \frac{3}{13} \frac{13-3-7}{13} \frac{5}{13} = 0.55 \cdot \frac{3 \cdot 3 \cdot 5}{13^3} = \frac{0.55 \cdot 45}{13^3} \\
P[C = \mathbf{B}|(\texttt{good}, \texttt{bad}, \texttt{okay})] &\propto P[C = \texttt{bad}]\ P[F_1 = \texttt{good}|\mathbf{B}]\ P[F_2 = \texttt{bad}|\mathbf{B}]\ P[F_3 = \texttt{okay}|\mathbf{B}] \\
&= 0.45 \cdot \frac{5}{11} \frac{11-3-5}{11} \frac{3}{11} = 0.45 \cdot \frac{5 \cdot 3 \cdot 3}{11^3} = \frac{0.45 \cdot 45}{11^3}
\end{aligned}$$

So,

$$\frac{P[C = \mathbf{A}|(\texttt{good}, \texttt{bad}, \texttt{okay})]}{P[C = \mathbf{B}|(\texttt{good}, \texttt{bad}, \texttt{okay})]} = \frac{0.55 \cdot 45 \cdot 11^3}{0.45 \cdot 45 \cdot 13^3} = \frac{11^4}{9 \cdot 13^3} = \frac{14641}{19773} = 0.7405$$

And therefore,

$$P[C = \mathbf{A}|(\texttt{good}, \texttt{bad}, \texttt{okay})] \approx \frac{0.7405}{1 + 0.7405} \approx\ 0.425$$

Yes, this is too much arithmetic for a closed calculator exam.

Solution: Partial credit solution. This is a solution in which the effect of the prior is completely neglected.
If someone explicitly wrote that they were using an improper BetaDist$(0,0)$ prior, I might have allowed that, but no one did. So I have assumed students with an answer like the one below simply neglected the notion of priors.
Let's start by writing the defining equation for Naïve Bayes

$$P[C|F_1,F_2,F_3] \propto P[C]\ P[F_1,F_2,F_3|C] \approx\ P[C]\ P[F_1|C]\ P[F_2|C]\ P[F_3|C]$$

So the parameters needed are $P[C]$, and the three $P[F|C]$ terms.
5 out of 9 data samples are of class **A**, so without pseudo-counts, $P[C=\mathbf{A}] = \frac{5}{9}$, and $P[C=\mathbf{B}] = \frac{4}{9}$.
Likewise the estimates of $P[F|C]$ are also just taken directly from counts in the dataset.

$$\text{no-pseudocount estimate } P[F_i|C_j] = \frac{N(F_i,C_j)}{N(C_j)}$$

Where $N(\texttt{X})$ denotes the count of `X` in the data. The following table summarizes those counts.

	feature F_1				feature F_2				feature F_3			
	class **A**		class **B**		class **A**		class **B**		class **A**		class **B**	
	`good`	`okay`	`good`	`okay`	`good`	`okay`	`good`	`okay`	`good`	`okay`	`good`	`okay`
$N(F_i\|C_j)$	1	1	2	2	1	3	1	2	3	2	0	1
$P[F_i\|C_j]$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{0}{4}$	$\frac{1}{4}$

Where $N[\texttt{X}]$ denotes the number of `X` observed in the data.
And $P[\texttt{bad}|C]$ can be computed as $1 - P[\texttt{good}|C] - P[\texttt{okay}|C]$.
Here I list an example to demonstrate that we have defined enough parameters.

$$\begin{aligned}
P[C=\mathbf{A}|(\texttt{good},\texttt{bad},\texttt{okay})] &\propto P[C=\texttt{good}]\ P[F_1=\texttt{good}|\mathbf{A}]\ P[F_2=\texttt{bad}|\mathbf{A}]\ P[F_3=\texttt{okay}|\mathbf{A}] \\
&= \frac{5}{9}\cdot\frac{1}{5}\frac{5-1-3}{5}\frac{2}{5} = \frac{5}{9}\cdot\frac{1\cdot1\cdot2}{5^3} = \frac{2}{9\cdot5^2}
\end{aligned}$$

$$\begin{aligned}
P[C=\mathbf{B}|(\texttt{good},\texttt{bad},\texttt{okay})] &\propto P[C=\texttt{bad}]\ P[F_1=\texttt{good}|\mathbf{B}]\ P[F_2=\texttt{bad}|\mathbf{B}]\ P[F_3=\texttt{okay}|\mathbf{B}] \\
&= \frac{4}{9}\cdot\frac{2}{4}\frac{4-1-2}{4}\frac{1}{4} = \frac{4}{9}\cdot\frac{2\cdot1\cdot1}{4^3} = \frac{2}{9\cdot4^2} =
\end{aligned}$$

So,

$$\frac{P[C=\mathbf{A}|(\texttt{good},\texttt{bad},\texttt{okay})]}{P[C=\mathbf{B}|(\texttt{good},\texttt{bad},\texttt{okay})]} = \frac{\frac{2}{9\cdot5^2}}{\frac{2}{9\cdot4^2}} = \frac{4^2}{5^2} = \frac{16}{25} = 0.64$$

And therefore,

$$P[C=\mathbf{A}|(\texttt{good},\texttt{bad},\texttt{okay})] = \frac{\frac{16}{25}}{1+\frac{16}{25}} = \frac{\frac{16}{25}}{\frac{25+16}{25}} = \frac{16}{41} \approx\ 0.390$$