Practice Final Exam
EECS 545: Machine Learning
Winter, 2016

Name:

UM uniqname:

- **Closed book. Three sheets of paper of notes are allowed. No computers, cell phones or calculators.**
  - Showing your work makes partial credit possible.
    If you write nothing at all, it's hard to justify any score but zero.
  - Feel free to use the backs of the sheets for scratch paper.
  - Write clearly. If we can't read your writing, it will be marked wrong.
- This course operates under the rules of the College of Engineering Honor Code. Your signature endorses the pledge below. **After** you finish your exam, please sign below:
  I have neither given nor received aid on this examination, nor have I concealed any violations of the Honor Code.

Problem 1 (True/False). Are the following statements true or false? (No need for explanations unless you feel the question is ambiguous and want to justify your answer).

1. Directed edges in a graphical model represent causal relationships.

   False

2. Let X and Y be two random variables on a discrete space $\{1, \ldots, n\}$. If X and Y are independent, then their mutual information $I(X;Y)$ is zero.

   True

3. Increasing the number of layers in a neural network always decrease the classification error of test data.

   False

4. The BackPropagation learning algorithm guarantees to find the globally optimal solution.

   False. It guarantees to find at least a local minimum of the error function.

5. Imagine we have a random variable X, and we drawn n IID samples $X_1, \ldots, X_n$. Using these samples we have two different estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ for mean of the distribution. It is necessarily the case that if $\hat{\mu}_1$ has higher bias than $\hat{\mu}_2$ then $\hat{\mu}_1$ has lower variance than $\hat{\mu}_2$.

   False

6. The k-means clustering algorithm minimizes a certain objective function. This objective function depends on the number of clusters k. It is possible that the global minimum of this objective can increase if we increase k.

   False

7. Suppose we run k-means on data generated from some Gaussian mixture model. If (by some luck) the cluster means obtained by the k-means algorithm are exactly the true cluster means of the model, then it is necessarily the case the the cluster assignments will also be the same.

False. It is simple to construct a Gaussian mixture model such that a point may be closer to one mixture component center but have a higher probability in another mixture component.

8. We are able to sample examples $\mathbf{x} \in \mathbb{R}^d$ from a multivariate Gaussian distribution. Then there is a linear transformation (matrix) $A \in \mathbb{R}^{d \times d}$ such that the transformed data $\mathbf{y} = A\mathbf{x}$ has the property that the entries in $\mathbf{y}$ are all mutually independent random variables.

   True

9. You are given a training dataset with features $A_1, ..., A_m$ and labeled example $x^{(1)}, x^{(2)}, ..., x^{(n)}$ and you build a decision tree D_1 selecting which feature to split on at each iteration based on information gain. You then take one of the labeled examples (say $x^{(2)}$), add a copy of it to the training set, and rerun your learning algorithm to create another decision tree D_2. Then D_1 and D_2 will necessarily be identical decision trees.

   False. Adding an instance can change the information gain calculated for each feature and so can change which attribute is chosen. In addition, for a leaf node, the plurality of the classification can change.

Problem 2 (Clustering). Consider the EM algorithm for the Gaussian mixture model with K mixture components for variable $\mathbf{x}$ with pdf

$$f(\mathbf{x}) = \sum_{k=1}^{K} w_k \phi(\mathbf{x}; \mu_k, \sigma^2 I)$$

The updates for the estimates $\hat{w}_k$ and $\hat{\mu}_k$ are as follows, given N iid draws from $f(\mathbf{x})$:

$$\gamma_{i,k} = \frac{w_k \phi(\mathbf{x}_i; \mu_k, \sigma^2 I)}{\sum_\ell w_\ell \phi(\mathbf{x}_i; \mu_\ell, \sigma^2 I)}$$
$$\hat{w}_k = \frac{1}{N} \sum_{i=1}^{N} \gamma_{i,k}$$
$$\hat{\mu}_k = \frac{\sum_{i=1}^{N} \gamma_{i,k} \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_{i,k}}$$

Show that if $\sigma^2 \to 0$, this algorithm becomes the k-means algorithm.

As $\sigma^2 \to 0$, all mixture component probabilities will be very small for any point, but the mixture component with the closest center to that point will dominate (assuming the point is not equidistant from more than one center, which has almost 0 probability), so

$$\gamma_{i,k} = \begin{cases} 1 & : \ k = \underset{\ell}{\operatorname{argmin}} \|\mathbf{x}_i - \mu_\ell\|_2^2 \\ 0 & : \ \text{otherwise} \end{cases}$$

So computing $\gamma_{i,k}$ becomes computing the cluster assignments. Then

$$\hat{\mu}_k = \frac{1}{\sum_{i=1}^{N} \gamma_{i,k}} \sum_{i=1}^{N} \gamma_{i,k} \mathbf{x}_i = \frac{1}{N_k} \sum_{\mathbf{x}_i : \gamma_{i,k} = 1} \mathbf{x}_i$$

which are the cluster means, where N_k is the number of points assigned to cluster k.

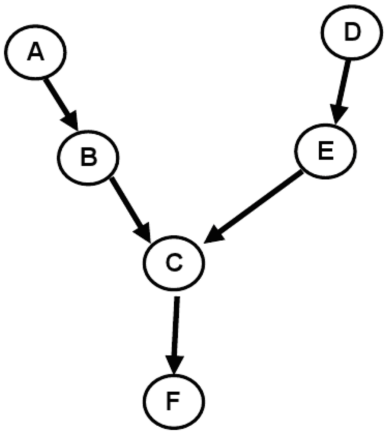


Figure 1: Bayes Net

Problem 3 (Bayesian Network). Given Figure 1, answer the following questions.

(A) Which of the following statements are true?

(a) $A \perp F$

(b) $A \perp F \mid C$ – True

(c) $A \perp D$ – True

(d) $A \perp D \mid C$

(e) $A \perp B$

(f) $A \perp B \mid C$

(g) $B \perp A$

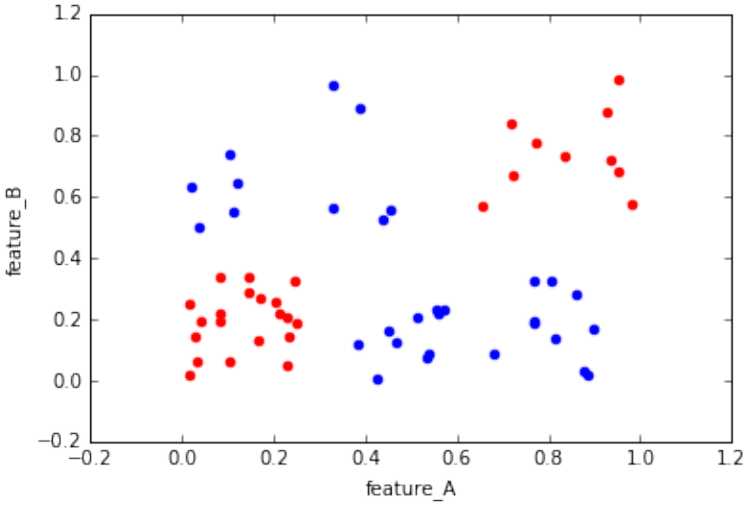
(h) $B \perp A \mid D$

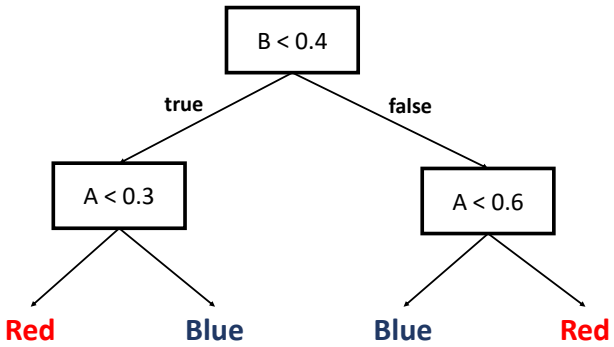
(i) $B \perp C \mid D$

(B) If $W \perp\!\!\!\perp X|Z$ and $X \perp\!\!\!\perp Y|Z$ for some distinct variables W, X, Y, Z, can you say $W \perp\!\!\!\perp Y|Z$? If so, show why. If not, find a counterexample from the graph above.

No. $A \perp\!\!\!\perp F|B$ and $D \perp\!\!\!\perp A|B$ but D and F are not independent given B.

Problem 4 (Decision Tree). Below we have input data with two features, `featureA` and `featureB`. Describe a decision tree that lets us perfectly classify this data with a depth-two decision tree.





Problem 5 (Neural Networks). Draw a 2 layer neural network with threshold units in the hidden layer and the output layer that will implement the boolean function specified in the table below (the first 3 columns are the inputs and the rightmost column is the output. Specify **all the weights** clearly).

0	0	0	0
1	0	0	0
0	1	0	0
1	1	0	1
0	0	1	0
1	0	1	0
0	1	1	1
1	1	1	0

Denote inputs from left to right as a_1, a_2, a_3 and output as o, then

$$o = (a_1 a_2 \neg a_3) \vee (\neg a_1 a_2 a_3)$$

The neural network is shown in Figure 2.

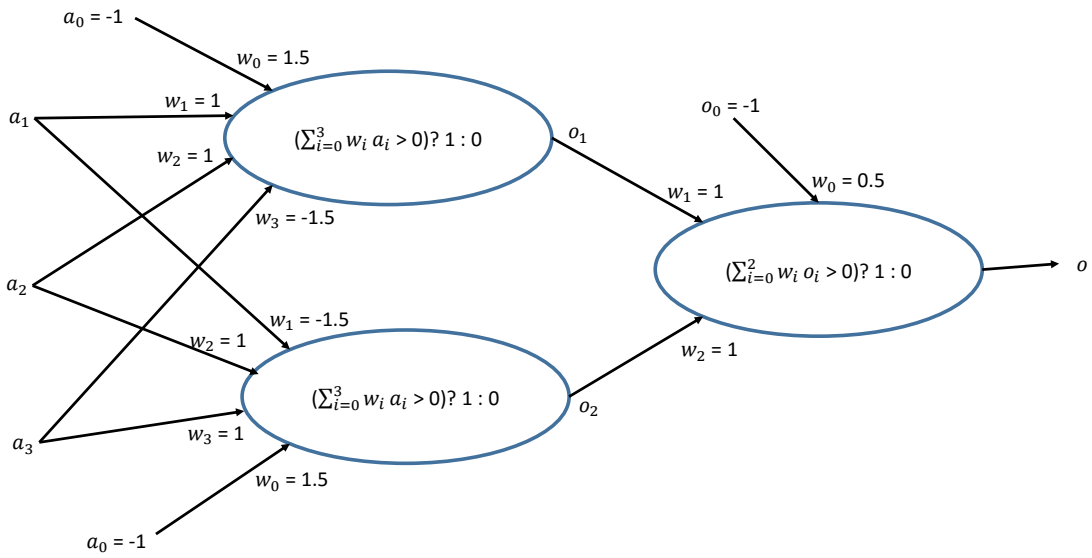


Figure 2: Neural network

Problem 6 (PCA). Prove that the matrix $U = [\mathbf{u}_1 \dots \mathbf{u}_K] \in \mathbb{R}^{D\times K}$, $K \leq D$, where $\mathbf{u}_k$ is the k^{th} principal component of $X \in \mathbb{R}^{D\times N}$, $N \geq D$, is a solution to the problem

$$\min_{A\in\mathbb{R}^{D\times K}: A^TA=I_K} \|X - AA^TX\|_F$$

In other words, prove that PCA provides us with a subspace $S \subseteq \mathbb{R}^d$ of dimension K that best preserves the data when the columns of X are projected onto S.

If we manipulate the objective function a bit, we see that

$$\begin{aligned}
\|X - AA^TX\|_F^2 &= \text{tr}(XX^T) - 2\text{tr}(AA^TXX^T) + \text{tr}(X^TAA^TAA^TX) \\
&= \text{tr}(XX^T) - 2\text{tr}(A^TXX^TA) + \text{tr}(X^TAA^TX) \\
&= \text{tr}(XX^T) - 2\text{tr}(A^TXX^TA) + \text{tr}(A^TXX^TA) \\
&= \text{trace}(XX^T) - \sum_{k=1}^{K} a_k^TXX^Ta_k
\end{aligned}$$

Then minimizing the objective is maximizing

$$\sum_{k=1}^{K} a_k^TXX^Ta_k$$

which we proved in homework is solved by PCA.

Problem 7 (EM). Consider a probabilistic model with random variables $z \in \{0, 1\}$, $\epsilon \in \mathbb{R}$, and $x \in \mathbb{R}$. The random variables z and ϵ are independent. The joint distribution of all three variables is as follows:

$$z \sim Bernoulli(\phi)$$
$$\epsilon \sim Normal(\mu, \sigma^2)$$
$$x = z + \epsilon$$

The parameters of this model are $\phi \in \mathbb{R}$, $\mu \in \mathbb{R}$, and $\sigma \in \mathbb{R}$.

(a) Write down a simple, one-line description of what the marginal distribution of x looks like.

Mixture of two Gaussians at μ and $\mu + 1$.

(b) Suppose we have a training set $\{(z^{(1)}, \epsilon^{(1)}, x^{(1)}), ..., (z^{(N)}, \epsilon^{(N)}, x^{(N)})\}$, where all three variables z, ϵ, x are observed. Write down the log-likelihood of the variables, and derive the maximum likelihood estimates of the models parameters.

$$p(z, \epsilon, x; \phi, \mu, \sigma) = p(x|z, \epsilon; \phi, \mu, \sigma)p(z, \epsilon; \phi, \mu, \sigma) = p(z, \epsilon; \phi, \mu, \sigma) = p(z; \phi)p(\epsilon; \mu, \sigma)$$

The first step uses the fact that $x = z + \epsilon$. The last step follows from the fact that z and ϵ are independent. Now,

$$l(\phi, \mu\sigma) = \sum_{i=1}^{N} (z^{(i)} log\phi + (1 - z^{(i)}) log(1 - \phi) + log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(\epsilon^{(i)} - \mu)^2}{2\sigma^2})$$

Now take the derivative of $l(\phi, \mu\sigma)$ with respect to ϕ and set to zero to get ML estimates for ϕ.

$$\frac{\partial l}{\partial \phi} = \sum_{i=1}^{N} (\frac{z^{(i)}}{\phi} - \frac{(1 - z^{(i)})}{1 - \phi}) = 0$$

$$\phi = \frac{1}{N} \sum_{i=1}^{N} z^{(i)}$$

Similarly, set $\frac{\partial l}{\partial \mu}$ and $\frac{\partial l}{\partial \sigma}$ to get

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \epsilon^{(i)}$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (\epsilon^{(i)} - \mu)^2$$

(c) Now, suppose z and ϵ are latent (unobserved) random variables. Our training set is therefore of the form $\{x^{(1)}, ..., x^{(N)}\}$. Write down the log-likelihood of the variables, and derive an EM algorithm to maximize the log-likelihood. Clearly indicate what are the E-step and the M-step.

$$p(x;\phi,\mu,\sigma) = p(x|z=1;\mu,\sigma)p(z=1;\phi) + p(x|z=0;\mu,\sigma)p(z=0;\phi)$$
$$= p(\epsilon = x-1;\mu,\sigma)\phi + p(\epsilon = x;\mu,\sigma)(1-\phi)$$
$$l(\mu,\phi,\sigma) = \sum_{i=1}^{N} log\frac{1}{\sqrt{2\pi}\sigma} + log(\phi \exp(-\frac{(x^{(i)}-1-\mu)^2}{2\sigma^2}) + (1-\phi)\exp(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}))$$

We treat z as a latent variable.
(E-Step)

$$Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}) = \frac{p(x^{(i)}|z)p(z^{(i)})}{p(x)}$$

$$Q_i(1) = \frac{\phi \exp(-\frac{(x^{(i)}-1-\mu)^2}{2\sigma^2})}{\phi \exp(-\frac{(x^{(i)}-1-\mu)^2}{2\sigma^2}) + (1-\phi)\exp(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2})}$$

(M-step)

$$\arg\max \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x,z)}{Q_i(z^{(i)})}$$

$$\phi = \frac{1}{N}\sum_{i=1}^{N} Q_i(1)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}((x-1)Q_i(1) + x(1-Q_i(1)))$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Q_i(1)(x-1-\mu)^2 + (1-Q_i(1))(x-\mu)^2)$$

Problem 8 (Bias/Variance). Suppose we are given 99 iid draws $S = (X_1, X_2, \ldots, X_{100})$ from a Bernoulli distribution and want to estimate the parameter p of the distribution. Compare the following two estimators:

$$\hat{p}_A = \text{mean}(S)$$
$$\hat{p}_B = \text{median}(S)$$

If $p \in [0.8, 0.9]$, which of the following is true?

(A) $\text{bias}^2(\hat{p}_A) > \text{bias}^2(\hat{p}_B)$, $\quad \text{variance}(\hat{p}_A) > \text{variance}(\hat{p}_B)$

(B) $\text{bias}^2(\hat{p}_A) > \text{bias}^2(\hat{p}_B)$, $\quad \text{variance}(\hat{p}_A) < \text{variance}(\hat{p}_B)$

(C) $\text{bias}^2(\hat{p}_A) < \text{bias}^2(\hat{p}_B)$, $\quad \text{variance}(\hat{p}_A) > \text{variance}(\hat{p}_B)$

(D) $\text{bias}^2(\hat{p}_A) < \text{bias}^2(\hat{p}_B)$, $\quad \text{variance}(\hat{p}_A) < \text{variance}(\hat{p}_B)$

Let's compute the expectations of our predictors. Let $X_{(50)}$ denote the 50^{th} order statistic of the sample, which is the median:

$$\mathbb{E}[\hat{p}_A] = \mathbb{E}\left[\frac{1}{99}\sum_{i=1}^{99} X_i\right] = \frac{1}{99}99\mathbb{E}[X_i] = p$$

$$\mathbb{E}[\hat{p}_B] = \mathbb{E}[X_{(50)}] = P(X_{(50)} = 1) = P(50 \text{ or more } X_i\text{'s are } 1) = \sum_{i=50}^{99} \binom{99}{i} p^i (1-p)^{99-i}$$

For $p \in [0.8, 0.9]$, $\mathbb{E}[\hat{p}_B] \approx 1$. Clearly $\hat{p}_B$ has a higher bias since $\hat{p}_A$ is unbiased. Now let us compute the variance:

$$\text{variance}(\hat{p}_A) = \frac{1}{99^2}\sum_{i=1}^{99} \text{variance}(X_i) = \frac{p(1-p)}{99}$$

$$\text{variance}(\hat{p}_B) = \text{variance}(X_{(50)}) = \mathbb{E}[X_{(50)}](1 - \mathbb{E}[X_{(50)}])$$

For $p \in [0.8, 0.9]$, $\text{variance}(\hat{p}_B) \approx 0$, and is much smaller than $\text{variance}(\hat{p}_A)$. See Figure 3 for plots of the bias squared and variance as functions of p.

Answer: (C)

When $p = 0.5$, which estimator has higher variance?

(A) $\hat{p}_A$

(B) $\hat{p}_B$

When $p = 0.5$, $P(X_{(50)} = 1) = P(X_{(50)} = 0) = 0.5$, so $\text{variance}(\hat{p}_A) = 0.25/99 < \text{variance}(\hat{p}_B) = 0.25$.

Answer: (B)

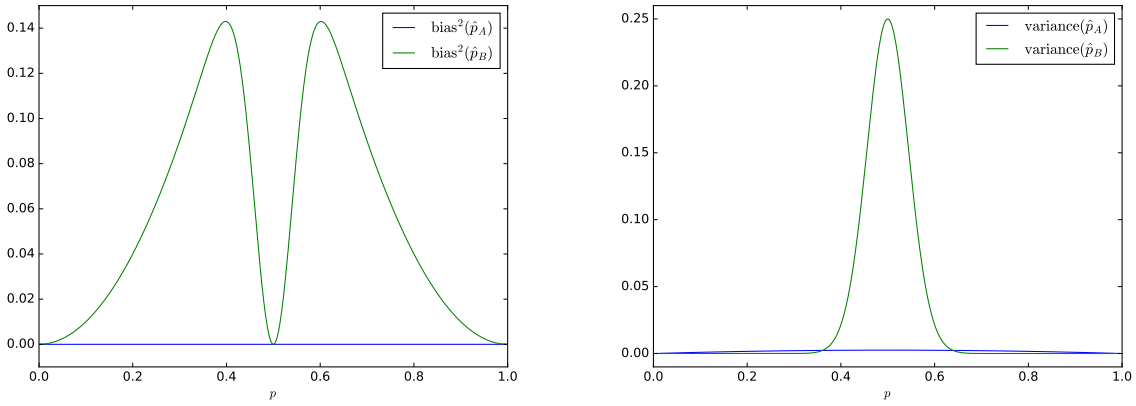


Figure 3: *Left:* Bias squared as a function of p. *Right:* Variance as a function of p.

Problem 9 (HMM). Suppose you have the following data on the quality of lectures from a Professor (where G stands for "good" and B stands for "bad")

$$GGBGBBBBGGBGGB$$

(A) Write down a first-order Markov model (and show the calculations; you can leave quantities as fractions). If there is not enough data to compute some part of the model, make some assumption about it and write it down.

	G	B
G	$\frac{3}{13}$	$\frac{4}{13}$
B	$\frac{3}{13}$	$\frac{3}{13}$

(B) Write down a second-order Markov model (and show the calculations; you can leave quantities as fractions). If there is not enough data to compute some part of the model, make some assumption about it and write it down.

	G	B
GG	0	$\frac{1}{4}$
GB	$\frac{1}{6}$	$\frac{1}{12}$
BG	$\frac{1}{6}$	$\frac{1}{12}$
BB	$\frac{1}{12}$	$\frac{1}{6}$

(C) What is the probability of the first two lectures being GB if we are given the following HMM model in Figure 4. Assume that the Professor is equally likely to start in the three

states (super-busy, not-busy, and busy). Show the formula that can be used to compute the probability of a sequence of observations from an HMM and then show the calculations needed to compute the answer to this specific question.

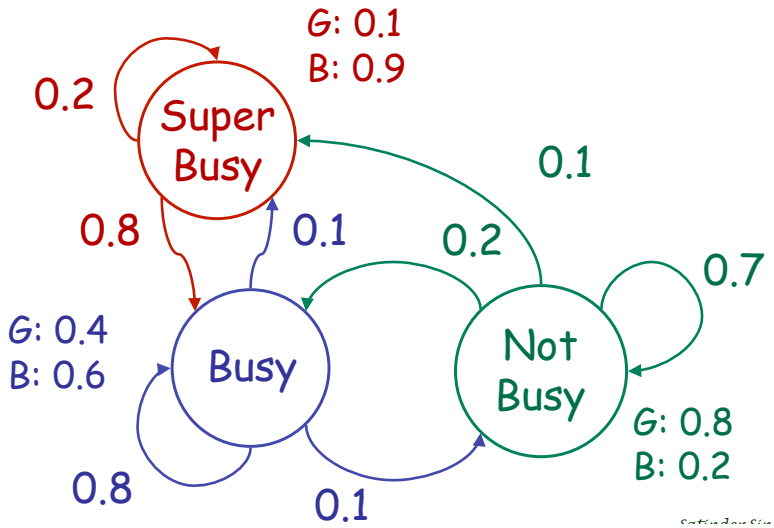


Figure 4: HMM

	SuperBusy	Busy	NotBusy
SuperBusy	0.2	0.8	0
Busy	0.1	0.8	0.1
NotBusy	0.1	0.2	0.7

Table 1: Transition probability

	G	B
SuperBusy	0.1	0.9
Busy	0.4	0.6
NotBusy	0.8	0.2

Table 2: Observation probability

Initial state probability:

$$P(SuperBusy) = P(Busy) = P(NotBusy) = \frac{1}{3}$$

$$P(GB) = P(SuperBusy)P(G|SuperBusy)[P(B|SuperBusy)P(SuperBusy|SuperBusy)$$
$$+P(B|NotBusy)P(NotBusy|SuperBusy) + P(B|Busy)P(Busy|SuperBusy)]$$
$$+P(Busy)P(G|Busy)[P(B|SuperBusy)P(SuperBusy|Busy)$$
$$+P(B|NotBusy)P(NotBusy|Busy) + P(B|Busy)P(Busy|Busy)]$$
$$+P(NotBusy)P(G|NotBusy)[P(B|SuperBusy)P(SuperBusy|NotBusy)$$
$$+P(B|NotBusy)P(NotBusy|NotBusy) + P(B|Busy)P(Busy|NotBusy)]$$