Exam instructions (please read carefully):

- **Exam structure.**
  - For 1000 level credit, you should attempt any 4 of the 6 problems. In this case, each of your solutions will be graded out of a maximum of 25 points. If you attempt more than 4, then the TAs will arbitrarily choose 4 of your solutions to grade. Note that the TAs may choose your 4 worst solutions.
  - For 2000 level credit, you should attempt all 6 problems. In this case, each of your solutions will be graded out a maximum of 17 points (yes - you can get 102!).

- **Allowed resources.** You are allowed to consult the course web site, your notes, and the course textbook. You are not allowed to use other resources, such as other books, other web sites, or other people.

- **Non-collaboration.** No collaboration is permitted on this exam. It is trusted that you will not discuss this exam or related course material with any other person (classmate or otherwise). You must abide by the Brown University Academic Code concerning examinations, quizzes, and tests (see http://goo.gl/mQtfSa).

- **Questions.** You may consult only with the CS155 instructor or TA for clarification of possible textual ambiguities. If you need such clarification, mail cs155tas@lists.brown.edu with your question. We will post clarifications to the class email list Do not make any posts to the class email list.

- **Handing in.** Print your solutions. Then print this instruction sheet, sign and write your name below, and attach the filled out sheet to the front of your packet. Handins must be submitted to the Computer Science Department receptionist on the 4th floor of the CIT. Late handins will not be accepted without prior authorization from the CS155 instructor.

**PLEASE SIGN (failure to sign voids the exam):** I solemnly state that I have abided by the exam instructions stated above, including the tenets of the Brown University Academic Code concerning examinations, quizzes and tests.

<table>
<thead>
<tr>
<th>Signature:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name (please print):</td>
</tr>
</tbody>
</table>

For use by graders

<table>
<thead>
<tr>
<th>Problem 1</th>
<th>Problem 2</th>
<th>Problem 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem 4</td>
<td>Problem 5</td>
<td>Problem 6</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1
Problem 1
Problem 4.20 from the textbook.

Problem 2
Dr. Smart is studying a new mysterious disease. He has identified a set \( D \) of \( n \) Boolean symptoms, such that each patient has some subset of this set, but no patient has more than \( d \) symptoms, for some fixed \( 0 < d \leq n \). Dr. Smart knows that there must be a subset \( C \subset D \) of symptoms that all patients have, and he wants to approximately learn this set while checking a minimum number of random patients.

(a) Formulate this question as a PAC learning problem.

(b) Define a range space associated with this learning problem.

(c) Prove that the VC-dimension of the range space is bounded by \( d \).

Problem 3
(a) Given two range spaces \( (X, R_1) \) and \( (X, R_2) \), each with VC-dimension \( d \). We define a “union” range space \( (X, R_{1 \cup 2}) \) where
\[
R_{1 \cup 2} = \{ r_1 \cup r_2 | r_1 \in R_1 \text{ and } r_2 \in R_2 \}.
\]
Prove that the VC dimension of \( R_{1 \cup 2} \) is \( O(d) \).

(b) Environmental scientists divide the country into polluted and not polluted areas. They want to identify the 5 axis-aligned squares that most accurately represent the polluted areas. We assume the map of the country is drawn on a 2D plane, and the axis-aligned squares are chosen from this map. Now, let \( C = \{ S_1, \ldots, S_5 \} \) be a set of 5 axis-aligned squares (each \( S_i \) is an axis-aligned square). For a randomly chosen location (point) \( X \) on the map, we denote \( X \in C \) if \( X \in \bigcup_{i=1}^{5} S_i \). We also denote \( P \) as the set of all polluted locations on the map. Define the error of \( C \) as
\[
\text{err}(C) = \Pr(X \in P, X \notin C) + \Pr(X \notin P, X \in C),
\]
which measure how well \( C \) minimizes the coverage of non-polluted areas and maximizes the coverage of polluted areas. Let \( C^* \) be the set of the 5 squares with minimum error. The scientists can sample uniformly at random a location in the country and characterize it as polluted or not polluted. For a given \( \epsilon > 0 \), their goal is to find a set of 5 squares \( C' \) such that \( \text{err}(C') \leq \text{err}(C^*) + \epsilon \). How many samples do they need in order to find the best 5 axis-aligned squares with probability at least \( 1 - \delta \), for \( \delta \in (0,1) \)? You can give the sample complexity using big \( O(\cdot) \) notation.

Problem 4
Problem 12.18 from the textbook. \( [G_{n,N} \text{ is the sample space of all graphs with } n \text{ vertices and } N \text{ edges. All graphs in this space have equal probability.}] \) **Clarification:** There is a typo in the question. You should let \( X \) be the “number of isolated vertices”, not the “expected number of isolated vertices”.
Problem 5

Let $\Omega$ be the set of all subsets of size $n$ of the set $\{1, \ldots, 2n\}$. Consider the following Markov chain on $\Omega$:

1. $X_1$ is an arbitrary subset in $\Omega$.

2. To compute $X_{t+1}$ choose an element $a$ uniformly at random from $X_t$ and an element $b$ uniformly at random from $\{1, \ldots, 2n\} \setminus X_t$. With probability $1/2$, $X_{t+1} = (X_t \setminus \{a\}) \cup \{b\}$, else $X_{t+1} = X_t$.

(a) Prove that the Markov chain is ergodic with uniform stationary distribution.

(b) Use a coupling argument to show that the mixing time $\tau(\epsilon)$ of the chain is $O(n \log(n/\epsilon))$. [Hint: Let $Y_t$ be the second chain. In the coupling process match elements in $X_t \cap Y_t$ to elements in that set, and elements in $X_t \setminus Y_t$ to elements in $Y_t \setminus X_t$. Do the same for the complementing sets.]

Problem 6

Problem 10.10 from the textbook.