STAT 538 Final Exam
Monday March 17 2014, 10:30-12:20

Student name: ..........................................................

- 16 pages of notes allowed
- no other sources of information are allowed
- electronic devices are not allowed

- Do Well!

<table>
<thead>
<tr>
<th>Problem</th>
<th>Points</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem 1</td>
<td>4</td>
<td>(KL divergence)</td>
</tr>
<tr>
<td>Problem 2</td>
<td>2</td>
<td>(Max entropy)</td>
</tr>
<tr>
<td>Problem 3</td>
<td>4</td>
<td>(Linear C-SVM)</td>
</tr>
<tr>
<td>Problem 4</td>
<td>8.5</td>
<td>(Bayes net)</td>
</tr>
<tr>
<td>Problem 5</td>
<td>5</td>
<td>(Junction tree)</td>
</tr>
<tr>
<td>Bonus</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

Total 24 points
Extra credit 3 points
Problem 1 – KL divergence and simple probability sets

1.1 (KL divergence ball) Let $\mathcal{P}_\Omega$ be the set of all probability distributions on $\Omega = \{1, 2, \ldots, m\}$, $q \in \mathcal{P}_\Omega$ and $a \geq 0$. Denote by $B_a = \{p \in \mathcal{P}_\Omega \mid D(p||q) \leq a\}$ where $D$ is the KL divergence$^1$.

Is $B_a$ convex? Prove or disprove.

---

$^1$Use the natural logarithm $\ln$ in the definition of KL divergence, as opposed to base 2 logarithm.
1.2 Let \( D \) be a set of \( N \) independent samples from \( p \in \mathcal{P}_\Omega \) and denote by \( \hat{p} \) the empirical distribution produced by \( D \). In other words, \( \hat{p}_j = \frac{1}{N}(\# \text{ times } j \text{ observed in sample}) \). Write the expression of \( \frac{1}{N} p(D) \) as a function of \( p \) and \( \hat{p} \).

1.3 (Sanov’s theorem, simplified) Using the result in 1.2, show that

\[
p(D) \leq e^{-N D(\hat{p} \| p)}
\]
Problem 2 – Maximum Entropy

Consider the binary variables $X, Y \in \{0, 1\}$ and denote by $p = [p_{00} \ p_{01} \ p_{10} \ p_{11}]^T$ a distribution $P_{XY}$ over $\{0, 1\}^2$.

2.1 We want to find the distribution $p^*$ that maximizes the entropy $H(p)$ subject to the constraint

$$P_X(1) = 2P_Y(1),$$

where $P_X, P_Y$ represent respectively the marginals of $p$ w.r.t. to $X, Y$.

Formulate this problem as a convex optimization problem in standard form.

2.2 Denote the dual variable corresponding to constraint (1) by $\theta$. Find the general parametric form of the distribution that solves the maximum entropy problem defined in 2.1. You are not required to derive the expression for $Z(\theta)$ for this question.
Extra credit: Show that the entries in $p^*$ obey a geometric progression. 

More credit for shorter proof.
Problem 3 – Linear C-SVM

Consider the linear SVM problem with slack variables for a data set \( \mathcal{D} = \{(x^i, y^i), i = 1 : N\} \). Assume that the data set \( \mathcal{D} \) is linearly separable.

\[
\min_{w, b, \xi} \quad \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i
\]
\[
\text{s.t.} \quad y^i(w^T x^i - b) \geq 1 - \xi_i \quad \text{for } i = 1 : N
\]
\[
\xi_i \geq 0 \quad \text{for } i = 1 : N
\]

In the above, \( C \geq 0 \) is a regularization parameter. We will study the dependence of the solution of (2) on \( C \), for a given \( \mathcal{D} \) as above.

Denote by \((w^*_C, b^*_C)\) the classifier parameters obtained from (2) with regularization parameter \( C \), and by \((w^*, b^*)\) the classifier parameter obtained when \( C = \infty \) (that is, without the use of slack variables).

3.1 Prove that there exists a \( C^* > 0 \) so that \((w^*_C, b^*_C) = (w^*, b^*)\) for all \( C \geq C^* \).
3.2 Prove that $||w_c^*||^2 \leq ||w^*||^2$ whenever $C < C^*$. 
Problem 4 – Bayesian Network

Consider the Bayesian Network

\[ \begin{array}{c}
\text{(No proofs required)} \\
\end{array} \]

4.1 In this graphical model, is \( B \perp C \)?

Yes No

4.2 In this graphical model, is \( B \perp C \mid D \)?

Yes No

4.3 In this graphical model, is \( B \perp C \mid A \)?

Yes No

4.4 Write the expression of a joint distribution \( P_{ABCD} \) which factorizes according to the DAG above.

\[ P_{ABCD} = \]
4.5 Assume all variables take values in \{0, 1\}. We observe the following sample of size \(N = 10\)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>#times observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

\[10 = N\]

Denote by \(\hat{P}_{ABCD}\) the empirical distribution defined by this sample (e.g. \(\hat{P}(0, 1, 1, 1) = 0.4\)) and by \(P_{ABCD}^{ML}, P_{A}^{ML}, \ldots\) the maximum likelihood estimate of the model (3) and of its parameters from this sample.

Give numerical answers to the following (no need to show work)

\[
P_{A}^{ML} =
\]
\[
P_{B|A}^{ML} =
\]
\[
P_{ABCD}^{ML}(0, 0, 0, 0) =
\]

4 points

4.6 Answer the following (no need to show work)

\[
\hat{P}_{A} = P_{A}^{ML} \quad \text{Yes} \quad \text{No}
\]
\[
\hat{P}_{ABCD} = P_{ABCD}^{ML} \quad \text{Yes} \quad \text{No}
\]
\[
\hat{P}_{B} = P_{B}^{ML} \quad \text{Yes} \quad \text{No}
\]
\[
B \perp C|A \text{ under } \hat{P} \quad \text{Yes} \quad \text{No}
\]
Problem 5 – Entropy of Junction Tree

5.1 Consider the MRF

Draw the Junction Tree corresponding to this MRF. Is this junction tree unique? (No need to show your work for this question)

5.2 Write the expression of a joint distribution $P_{ABCD}$ which factorizes according to the junction tree in 5.1. (No need to show your work for this question)
5.3 Derive the expression of the entropy of the distribution obtained in 5.2, showing that this expression involves only marginals of cliques and separators.

5.4 Extra credit Can you express the entropy in 5.3 as a function of entropies and mutual informations between pairs of variables?