STA314 Midterm Review
Midterm Review

1. A brief overview

2. Some past midterm questions
   (Note: midterm will be on Quercus and probably multiple choice)
• **Supervised learning and Unsupervised learning**

  * **Supervised learning**: have a collection of training examples labeled with the correct outputs

  * **Unsupervised learning**: have no labeled examples

• **Regression and Classification**

  * **Regression**: predicting a scalar-valued target

  * **Classification**: predicting a discrete-valued target
K-Nearest Neighbors

Idea: Classify a new input $\mathbf{x}$ based on its $k$ nearest neighbors in the training set

Decision boundary: the boundary between regions of input space assigned to different categories

Tradeoffs in choosing $k$: overfit / underfit

Pitfalls: curse of dimensionality, normalization, computational cost
- **Linear Regression**

  **Model**: a linear function of the features \( y = w^T x + b \)

  **Loss function**: squared error loss \( \mathcal{L}(y, t) = \frac{1}{2}(y - t)^2 \)

  **Cost function**: loss function averaged over all training examples

  **Vectorization**: advantages

  **Solving minimization problem**: direct solution / gradient descent \( w \leftarrow w - \alpha \frac{\partial J}{\partial w} \)

  **Feature mapping**: degree-M polynomial feature mapping
**Model Complexity and Generalization**

**Underfitting:** too simplistic to describe the data

**Overfitting:** too complex, fit training examples perfectly, but fails to generalize to unseen data

**Hyperparameter:** can’t include in the training procedure itself, tune it using a validation set

**Regularization:** \( J_{\text{reg}}(w) = J(w) + \lambda R(w) \), improve the generalization, L2 / L1 regularization

-Pattern Recognition and Machine Learning, Christopher Bishop.
• Linear Classification
  • Binary Linear Classification
    Model: \[ z = \mathbf{w}^\top \mathbf{x} \]
    \[ y = \begin{cases} 
    1 & \text{if } z \geq 0 \\
    0 & \text{if } z < 0 
  \end{cases} \]
  \[ y = \mathbb{1}[y \neq t] \]
  Geometry: input space, weight space

Loss function: 0-1 loss
\[ \mathcal{L}_{0-1}(y, t) = \begin{cases} 
    0 & \text{if } y = t \\
    1 & \text{if } y \neq t 
  \end{cases} \]

• Logistic Regression
  Model: \[ z = \mathbf{w}^\top \mathbf{x} \]
  \[ y = \sigma(z) \]
  Loss function: 0-1 loss
  \[ \mathcal{L}_{SE}(z, t) = \frac{1}{2}(z - t)^2 \]

  \[ \mathcal{L}_{SE}(y, t) = \frac{1}{2}(y - t)^2. \]

  \[ \mathcal{L}_{CE} = -t \log y - (1 - t) \log(1 - y) \]

• Softmax Regression
  Multi-class classification
  \[ y_k = \text{softmax}(z_1, \ldots, z_K)_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}} \]
  \[ z = \mathbf{W} \mathbf{x} \]
  \[ y = \text{softmax}(z) \]
  \[ \mathcal{L}_{CE} = -\mathbf{t}^\top \log \mathbf{y} \]
**Decision Trees**

**Model**: make predictions by splitting on features according to a tree structure

**Decision boundary**: made up of axis-aligned planes

**Entropy**: uncertainty inherent in the variable’s possible outcomes

$$H(Y) = - \sum_{y \in Y} p(y) \log_2 p(y)$$

joint entropy; conditional entropy; properties

**Information gain**: $IG(Y|X) = H(Y) - H(Y|X)$
measures the informativeness of a variable; used to choose a good split
Other topics to know

- Comparisons between different classifiers (KNN, logistic regression, decision trees, neural networks)
- Contrast the decision boundaries for different classifiers
7. [2pts] Consider the classification problem with the following dataset:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Your job is to find a linear classifier with weights $w_1$, $w_2$, $w_3$, and $b$ which correctly classifies all of these training examples. None of the examples should lie on the decision boundary.

(a) [1pt] Give the set of linear inequalities the weights and bias must satisfy.

(b) [1pt] Give a setting of the weights and bias that correctly classifies all the training examples. You don’t need to show your work, but it might help you get partial credit.
Many answers are possible. Here's one:

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$t = 1, w_1 x_1 + w_2 x_2 + w_3 x_3 + b > 0$

$t = 0, w_1 x_1 + w_2 x_2 + w_3 x_3 + b < 0$

\[
\begin{align*}
\{ & w_1 \cdot 0 + w_2 \cdot 0 + w_3 \cdot 0 + b > 0 \\
& w_1 \cdot 0 + w_2 \cdot 1 + w_3 \cdot 0 + b < 0 \\
& w_1 \cdot 0 + w_2 \cdot 1 + w_3 \cdot 1 + b > 0 \\
& w_1 \cdot 1 + w_2 \cdot 1 + w_3 \cdot 1 + b < 0 \\
\} \quad \Rightarrow \quad \\
\{ & b > 0 \quad \Rightarrow \quad b = 1 \\
& w_2 + b < 0 \quad \Rightarrow \quad w_1 = -2 \\
& w_2 + w_3 + b > 0 \quad \Rightarrow \quad w_2 = -2 \\
& w_1 + w_2 + w_3 + b < 0 \quad \Rightarrow \quad w_3 = 2
\end{align*}
\]
7. [2pts] Suppose binary-valued random variables $X$ and $Y$ have the following joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>$1/8$</td>
<td>$3/8$</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$2/8$</td>
<td>$2/8$</td>
</tr>
</tbody>
</table>

Determine the information gain $IG(Y|X)$. You may write your answer as a sum of logarithms.
Solution

<table>
<thead>
<tr>
<th></th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>1/8</td>
<td>3/8</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>2/8</td>
<td>2/8</td>
</tr>
</tbody>
</table>

$p(Y = 0) = p(X = 0, Y = 0) + p(X = 1, Y = 0) = \frac{3}{8}$

$p(Y = 1) = p(X = 0, Y = 1) + p(X = 1, Y = 1) = \frac{5}{8}$

$p(X = 0) = p(X = 0, Y = 0) + p(X = 0, Y = 1) = \frac{1}{2}$

$p(X = 1) = p(X = 1, Y = 0) + p(X = 1, Y = 1) = \frac{1}{2}$

$p(Y = 0 | X = 0) = \frac{p(Y = 0, X = 0)}{p(X = 0)} = \frac{p(Y = 0, X = 0)}{p(X = 0, Y = 0) + p(X = 0, Y = 1)} = \frac{1}{4}$

$IG(Y|X) = H(Y) - H(Y|X)$

$H(Y) = -\sum_y p(Y = y) \log_2 p(Y = y)$

$= -p(Y = 0) \log_2 p(Y = 0) - p(Y = 1) \log_2 p(Y = 1)$

$= -\frac{3}{8} \log_2 \frac{3}{8} - \frac{5}{8} \log_2 \frac{5}{8}$

$H(Y|X) = \sum_x p(X = x)H(Y|X = x)$

$= p(X = 0)H(Y|X = 0) + p(X = 1)H(Y|X = 1)$

$= \frac{1}{2}H(Y|X = 0) + \frac{1}{2}H(Y|X = 1)$

$H(Y|X = x) = -\sum_y p(y|x) \log_2 p(y|x)$

$H(Y|X = 0) = -p(Y = 0|X = 0) \log_2 p(Y = 0|X = 0)$

$= \frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4}$

$H(Y|X = 1) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$

We used: $p(y|x) = \frac{p(x,y)}{p(x)}$ and $p(x) = \sum_y p(x, y)$