Compare decision trees with kNN; what do they have in common; what are the main differences between the two approaches?

**Answer:**

Commonality: Decision trees and kNN are both supervised learning methods that assign a class to an object based on the features.

Main Differences: Decision trees define a hierarchy of rules in the form of trees and these rules are formed from training data. These trees give priority to more informative features. A decision tree could become very complex as the decision tree gets larger and requires pruning to improve its performance.

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learn Model (Tree)</td>
<td>Lazy Learner</td>
</tr>
<tr>
<td>Global Model</td>
<td>Local Model</td>
</tr>
<tr>
<td>Based on Attribute Order</td>
<td>Distance Based</td>
</tr>
<tr>
<td>Rectangle Decision Boundaries</td>
<td>Voronoi Convex Polygon Decision Boundary</td>
</tr>
</tbody>
</table>

Show that the VC dimension of the triangle hypothesis class is 7 in 2 dimensions (Hint: For best separation place the points equidistant on a circle). Generalizing your answer to the previous question, what is the VC dimension of a (non-intersecting) polygon of p points?

**Answer:**

As we can see in figure 2.4, for all possible labeling of seven points, we can draw a triangle to separate the positive and negative examples. We cannot do the same when there are eight points.
Intuitively, a p-sided polygon can contain at least p points. However, it cannot contain p + 1 (or more) points if there are more than 2p+1 and the points are placed in alternate positions (they are not next to each other). Therefore, the VC dimension of a (non-intersecting) polygon of p points is 2p + 1 (VC-dimension = 2p+1).

3. One major challenge when learning prediction and classification models is to avoid overfitting. What is overfitting? What factors contribute to overfitting? What is the generalization error? What is the challenge in determining the generalization error? Briefly describe one approach to determine the generalization error. 8-12 sentences!

**Answer:**
Overfitting: A model H is more complex than the underlying function f; when the model is too complex, the test errors are large although training errors are small.
Factors contribute to overfitting are:
1. Noise in the training examples that are not a part of the pattern in general data set.
2. Selected a hypothesis that is more complex than necessary.

... Generalization error is the error of the model on new data examples. The main challenge in determining the generalization error is that we don't actually have new data examples. The cross-validation approach is often used to estimate the generation error. It splits the training data into a number of subsets, one subset is used as the testing set and the rest of the data are used as the training set. The experiment repeats until all subsets have been used as the testing set. The generalization error is estimated by the average of testing errors.

4. Derive equation 2.17; the values for w_0 and w_1 which minimizes the squared prediction error!

**Answer:**

\[ E(w_1, w_0|x) = \frac{1}{N} \sum_{t=1}^{N} (r^t - (w_1 x^t + w_0))^2 \]

Partial derivatives of E with respect to w_0

\[ \frac{\partial E(w_1, w_0|x)}{\partial w_0} = \frac{1}{N} \sum_{t=1}^{N} \frac{\partial [r^t - (w_1 x^t + w_0)]^2}{\partial w_0} \]

\[ = \frac{1}{N} \sum_{t=1}^{N} (2r^t - 2w_1 x^t - 2w_0) \]

setting it equal to 0,
A lot of decision making systems use Bayes’ theorem relying on conditional independence assumptions—what are those assumptions exactly? Why are they made? What is the problem with making those assumptions? 3-6 sentences!

**Answer:**
The conditional independence assumptions of Bayes’ theorem are that every feature Fi is conditionally independent of every other feature Fj for i ≠ j for a given class.

It assumes the presence of a particular feature of a class is unrelated to the presence of any other feature. The assumption is made to simplify decision making by simplifying computations and by dramatically reducing knowledge acquisition cost. The problem of making those assumptions is that features are often correlated, and making this assumption in presence of correlation leads to errors in probability computations, and ultimately to making the wrong decision.
6. Assume we have a problem in which you have to choose between 3 decisions D1, D2, D3. The loss function is: \( \lambda_{11}=0, \lambda_{22}=0, \lambda_{33}=0, \lambda_{12}=1, \lambda_{13}=1, \lambda_{21}=1, \lambda_{31}=10, \lambda_{23}=1, \lambda_{32}=8 \); write the optimal decision rule! Decision rule! (\( \lambda_{ik} \) is the cost of choosing C\( i \) when the correct answer is C\( k \)). If you visualize the decision rule by Feb. 20, you get 50% extra credit; send your visualization and a brief description how you obtained it to Dr. Eick.

**Answer:**

\[
P(C_1) + P(C_2|x) + P(C_3|x) = 1
\]

\[
R(\alpha_i|x) = \sum_{k \neq i} \lambda_{ik} P(C_k|x)
\]

\[
R(\alpha_1|x) = \lambda_{12} P(C_2|x) + \lambda_{13} P(C_3|x) = P(C_2|x) + P(C_3|x) = 1 - P(C_1)
\]

\[
R(\alpha_2|x) = \lambda_{21} P(C_1|x) + \lambda_{23} P(C_3|x) = P(C_1|x) + P(C_3|x) = 1 - P(C_2)
\]

\[
R(\alpha_3|x) = \lambda_{31} P(C_1|x) + \lambda_{32} P(C_2|x) = 10P(C_1|x) + 8P(C_2|x)
\]

The optimal decision rule is:
Choose \( D_i \) if \( R(\alpha_i|x) < R(\alpha_k|x) \) for all \( k \neq i \)

Choose \( D_1 \) if
\[
R(\alpha_1|x) < R(\alpha_2|x)
\]
\[
1 - P(C_1) < 1 - P(C_2)
\]
\[
P(C_1) > P(C_2)
\]
and
\[
R(\alpha_1|x) < R(\alpha_3|x)
\]
\[
1 - P(C_1) < 10P(C_1|x) + 8P(C_2|x)
\]
\[
11P(C_1|x) + 8P(C_2|x) > 1
\]

Choose \( D_2 \) if
\[
R(\alpha_2|x) < R(\alpha_1|x)
\]
\[
1 - P(C_2) < 1 - P(C_1)
\]
\[
P(C_2) > P(C_1)
\]
and
\[
R(\alpha_2|x) < R(\alpha_3|x)
\]
\[
1 - P(C_2) < 10P(C_1|x) + 8P(C_2|x)
\]
\[
10P(C_1|x) + 9P(C_2|x) > 1
\]

Choose \( D_3 \) if
\[
R(\alpha_3|x) < R(\alpha_1|x)
\]
\[
10P(C_1|x) + 8P(C_2|x) < 1 - P(C_1)
\]
\[
11P(C_1|x) + 8P(C_2|x) < 1
\]
and
\[
R(\alpha_3|x) < R(\alpha_2|x)
\]
\[
10P(C_1|x) + 8P(C_2|x) < 1 - P(C_2)
\]
\[
10P(C_1|x) + 9P(C_2|x) < 1
\]
7. What does bias measure; what does variance measure? Assume we have a model with a high bias and a low variance—what does this mean? 3-4 sentences!

**Answer:**
Bias: measure the error between the estimator’s expected parameter and the real parameter. 
Variance: measures how much the estimator fluctuates around the expected value. 
A high bias and a low variance model is a simple model that is underfitting to the dataset.

8. Maximum likelihood, MAP, and the Bayesian approach all measure parameters of models. What are the main differences between the 3 approaches? 3-6 sentences!

**Answer:**
Maximum likelihood estimates the parameter by estimating the distribution that most likely resulted in the data. MAP and Bayesian approach both take into account the prior density of the parameter. MAP replaces the whole density with a single point to get rid of the evaluation of the integral, whereas the Bayesian approach uses an approximation method to evaluate the full integral.

9. Solve problem 4.10.6 in the textbook! Some hint will be given in Feb. 11 lecture! Assume we have a single attribute classification problem involving two classes C1 and C2 with the
following priors: \(P(C1)=0.6\) and \(P(C2)=0.4\). Give the decision rule\(^1\) assuming:
\[p(x|C1)\sim \mathcal{N}(0,4); \ p(x|C2)\sim \mathcal{N}(1,1)\]

**Answer:**

Normal Distribution

\[p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right]\]

\[p(x|C_1) = \frac{1}{\sqrt{2\pi}2} \exp\left[-\frac{x^2}{2*4}\right]\quad\text{and}\quad p(x|C_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-1)^2}{2}\right]\]

Using Bayes’ theorem, assume \(P(C_1) = 0.6\) and \(P(C_2) = 0.4\):

Equate the posterior probabilities to find their intersections:

\[P(x|C_1)P(C_1) = P(x|C_2)P(C_2)\]

\[\frac{1}{\sqrt{2\pi}2} \exp\left[-\frac{x^2}{2*4}\right] \cdot P(C_1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-1)^2}{2}\right] \cdot P(C_2)\]

Taking the log of both sides gives,

\[-\frac{1}{2} log2\pi - log2 - \frac{x^2}{8} + log P(C_1) = -\frac{1}{2} log2\pi - \frac{(x - 1)^2}{2} + log P(C_2)\]

\[-log2 - \frac{x^2}{8} + log P(C_1) = -\frac{(x - 1)^2}{2} + log P(C_2)\]

\[-log2 - \frac{x^2}{8} + log P(C_1) = -\frac{x^2}{2} + x - \frac{1}{2} + log P(C_2)\]

\[-8 log 2 - x^2 + 8 log P(C_1) = -4x^2 + 8x - 4 + 8 log P(C_2)\]

\[3x^2 - 8x + 4 - 8 log 2 + 8 log P(C_1) - 8 log P(C_2) = 0\]

\[3x^2 - 8x + 1.6985 = 0\]

Using the quadratic formula,

\[\text{point1} = \frac{8 - \sqrt{64 - 4 \cdot 3 \cdot 1.6985}}{6} = 0.2326\]

\[\text{point2} = \frac{8 + \sqrt{64 - 4 \cdot 3 \cdot 1.6985}}{6} = 2.4341\]

Decision rule:

- If \(x > 2.4341\), then choose C1
- else if \(x < 0.2326\), then choose C1
- else
  - choose C2.

\(^1\) Write the rule in the form: If \(x\ldots\) then\ldots else if \(\ldots\) else\ldots!
10. Assume we have a dataset with 3 attributes and the following covariance matrix $\Sigma$:

$$
\begin{bmatrix}
9 & 0 & 0 \\
0 & 4 & -1 \\
0 & -1 & 1
\end{bmatrix}
$$

a) What are the correlations between the three attributes?

b) Assume we construct 3-dimensional normal distribution for this dataset by using equation 5.7 assuming that the mean is $\mu = (0,0,0)$. Compute the probability of the three vectors: $(1,1,0)$, $(1,0,1)$ and $(0,1,1)$!

c) Compute the Mahalanobis distance between the vectors $(1,1,0)$, $(1,0,1)$ and $(0,1,1)$. Also compute the Mahalanobis distance between $(1,1,-1)$ and the three vectors $(1,0,0)$, $(0,1,0)$, $(0,0,-1)$. How do these results differ from using Euclidean distance? Try to explain why particular pairs of vectors are closer/further away from each other when using Mahalanobis distance. What advantages do you see in using Mahalanobis distance of Euclidean distance?

**Answer:**

Given a dataset with three attributes $X$, $Y$, and $Z$, the correlations are

$$
\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0}{3 \times 2} = 0
$$

$$
\rho_{YZ} = \frac{\sigma_{YZ}}{\sigma_Y \sigma_Z} = \frac{-1}{2 \times 1} = -0.5
$$

$$
\rho_{XZ} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = \frac{0}{3 \times 1} = 0
$$

b) Assuming that the mean is $\mu = (0,0,0)$, the probabilities of the three vectors, $(1,1,0)$, $(1,0,1)$ and $(0,1,1)$, are

$$
p((1,1,0)) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp \left[-\frac{1}{2} \left( \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right)^T \begin{pmatrix} 9 & 0 & 0 \\ 0 & 4 & -1 \\ 0 & -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right]
= 0.0098
$$

$$
p((1,0,1)) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp \left[-\frac{1}{2} \left( \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right)^T \begin{pmatrix} 9 & 0 & 0 \\ 0 & 4 & -1 \\ 0 & -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right]
= 0.0059
$$

$$
p((0,1,1)) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp \left[-\frac{1}{2} \left( \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right)^T \begin{pmatrix} 9 & 0 & 0 \\ 0 & 4 & -1 \\ 0 & -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right]
= 0.0038
$$

c) The Mahalanobis distances between the vectors $\bar{x} = (1,1,0), \bar{y} = (1,0,1)$, and $\bar{z} = (0,1,1)$ are

$$
d(\bar{x}, \bar{y}) = \sqrt{(\bar{x} - \bar{y})^T \Sigma^{-1} (\bar{x} - \bar{y})}
$$

$$
(\bar{x} - \bar{y}) = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}
$$
The Mahalanobis distances between \(\vec{x}, \vec{y}\) and the three vectors \(\vec{p}, \vec{q}, \vec{r}\) are:

\[
d(\vec{x}, \vec{y}) = \sqrt{[0, 1, -1] \times \begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 1/3 & 1/3 \\ 0 & 1/3 & 4/3 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}} = 1
\]

\[
d(\vec{y}, \vec{z}) = \sqrt{(\vec{y} - \vec{z})^T \Sigma^{-1} (\vec{y} - \vec{z})} = \sqrt{\begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 1/3 & 1/3 \\ 0 & 1/3 & 4/3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \frac{4}{9} = 2/3 \approx 0.667
\]

\[
d(\vec{x}, \vec{z}) = \sqrt{(\vec{x} - \vec{z})^T \Sigma^{-1} (\vec{x} - \vec{z})} = \sqrt{\begin{bmatrix} 1/9 & 0 & 0 \\ 0 & 1/3 & 1/3 \\ 0 & 1/3 & 4/3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}} = \frac{\sqrt{13}}{3} \approx 1.202
\]
For Euclidean distance, the distance for the 3 points to the point $(1,1,-1)$ are all $\sqrt{2}$. But as we seen from the computation, the Mahalanobis distance are different for all 3 points. The vector pairs $(\vec{x}, \vec{y})$ and $(\vec{w}, \vec{p})$ are closer together because attributes Y and Z are correlated. Vector pairs $(\vec{x}, \vec{z})$ and $(\vec{w}, \vec{q})$ are closer together because attribute X has a larger variance than the other attributes, which outweighs the impact of attribute Z having a smaller variance than the others. Vector pairs $(\vec{y}, \vec{z})$ and $(\vec{w}, \vec{r})$ have the smallest Mahalanobis distance because attributes X and Y have variances greater than one, which means the vectors are "closer" to the reference than if the variances were one.

The advantage of using Mahalanobis distance over the Euclidean distance is it that the Mahalanobis distance is normalized by the variance of the attribute and the correlation between attributes.