W4. Discrete Classification II - Decision Trees

Guang Cheng

University of California, Los Angeles

guangcheng@ucla.edu

Week 4

Outline

Issues in business decision making

- How to present complicated decision problems in a systematical, logic, and intuitive way?

Outline

Issues in business decision making

- How to present complicated decision problems in a systematical, logic, and intuitive way?
- How to make a choice when facing multiple alternatives?

Outline

Issues in business decision making

- How to present complicated decision problems in a systematical, logic, and intuitive way?
- How to make a choice when facing multiple alternatives?

Outline

Issues in business decision making

- How to present complicated decision problems in a systematical, logic, and intuitive way?
- How to make a choice when facing multiple alternatives?

Concepts and Methods

- Decision Tree

Outline

Issues in business decision making

- How to present complicated decision problems in a systematical, logic, and intuitive way?
- How to make a choice when facing multiple alternatives?

Concepts and Methods

- Decision Tree
- Expected monetary value (EMV)

Outline

Issues in business decision making

- How to present complicated decision problems in a systematical, logic, and intuitive way?
- How to make a choice when facing multiple alternatives?

Concepts and Methods

- Decision Tree
- Expected monetary value (EMV)
- Multi-stage decision problems

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:
  - Prices for the three properties are: \$25k (A), \$50k (B), \$100k (C).

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:
  - Prices for the three properties are: \$25k (A), \$50k (B), \$100k (C).
  - The commission rate is 4%.

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:
  - Prices for the three properties are: \$25k (A), \$50k (B), \$100k (C).
  - The commission rate is 4%.
- The client also requests:

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:
  - Prices for the three properties are: \$25k (A), \$50k (B), \$100k (C).
  - The commission rate is 4%.
- The client also requests:
  - Sell A first.

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:
  - Prices for the three properties are: \$25k (A), \$50k (B), \$100k (C).
  - The commission rate is 4%.
- The client also requests:
  - Sell A first.
  - If A is unsold within a month, the entire deal is off—no commission and no chance to sell the other properties.

Motivating case: Warren agency

- Mr. Warren operates a real estate agency. He is approached one day by a prospective client who wants to sell three properties (A, B, C).
- The client indicates:
  - Prices for the three properties are: \$25k (A), \$50k (B), \$100k (C).
  - The commission rate is 4%.
- The client also requests:
  - Sell A first.
  - If A is unsold within a month, the entire deal is off—no commission and no chance to sell the other properties.
  - If A is sold within a month, then Warren will get the commission for A and the option of stopping at this point or trying to sell either the B or C next under the same conditions (i.e., sell within a month or no commission on the second property and no chance to sell the third).

Warren's decision problem

After the client left, Warren estimated the costs, revenues, and the probability to sell each property within a month:

	Cost	Price	Commission (4%)	Probability
A	800	25000	1000	0.7
B	200	50000	2000	0.6
C	400	100000	4000	0.5

Discussion

- Will you accept the business if you were Warren?

Warren's decision problem

After the client left, Warren estimated the costs, revenues, and the probability to sell each property within a month:

	Cost	Price	Commission (4%)	Probability
A	800	25000	1000	0.7
B	200	50000	2000	0.6
C	400	100000	4000	0.5

Discussion

- Will you accept the business if you were Warren?
- If you decide to take it, and are able to sell A first, what would your next step be? Then the next

Warren's decision problem

After the client left, Warren estimated the costs, revenues, and the probability to sell each property within a month:

	Cost	Price	Commission (4%)	Probability
A	800	25000	1000	0.7
B	200	50000	2000	0.6
C	400	100000	4000	0.5

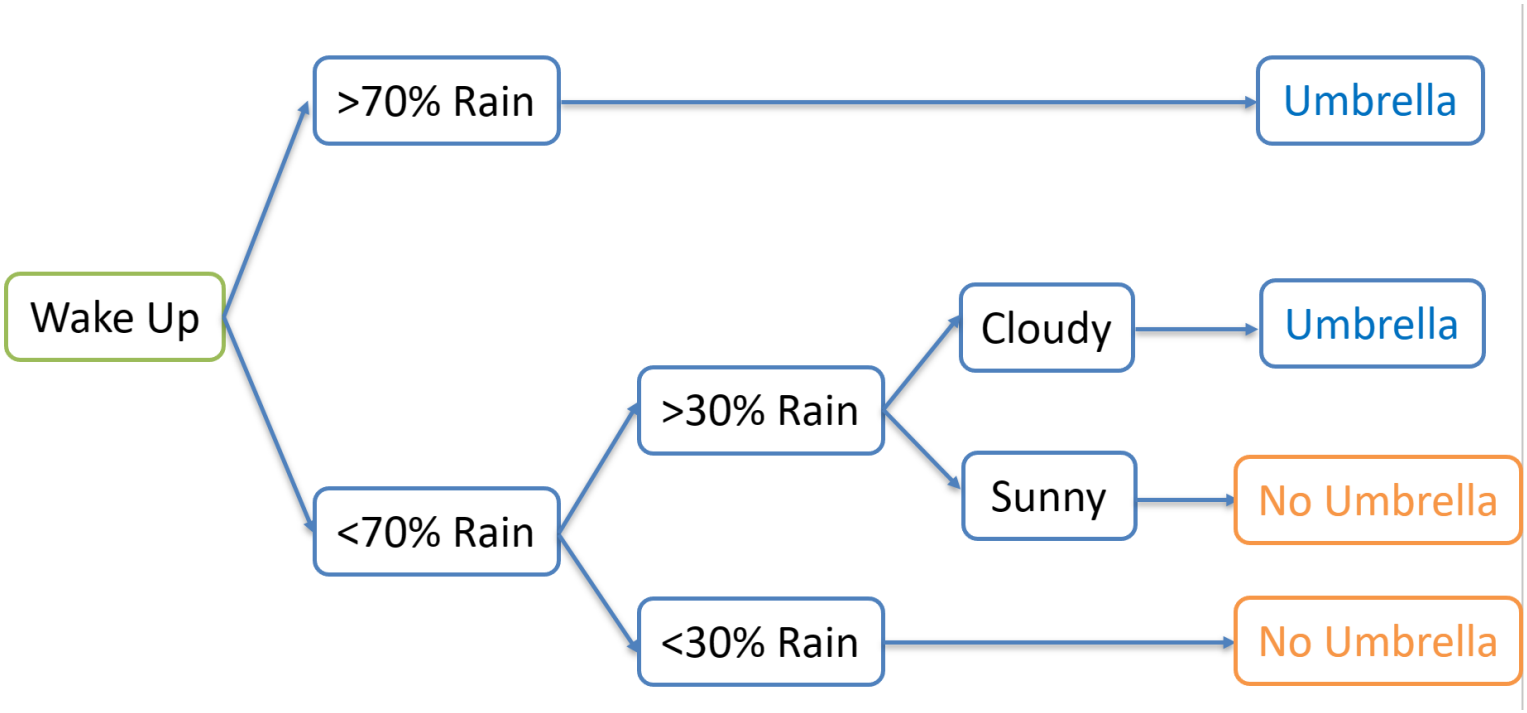
Discussion

- Will you accept the business if you were Warren?
- If you decide to take it, and are able to sell A first, what would your next step be? Then the next
- On what basis would you make your decisions?

Can we present a complex decision problem in a logical and intuitive way?

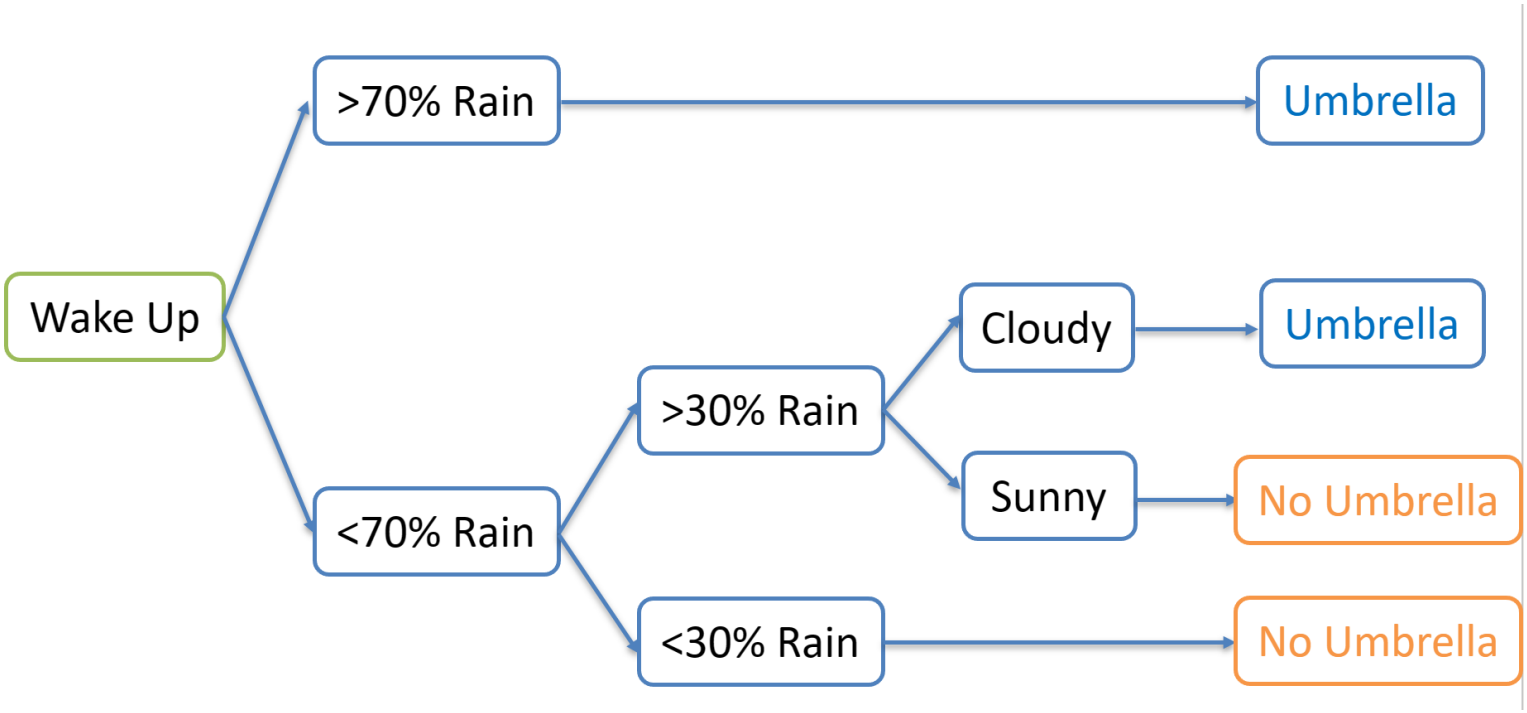
- Using decision tree !

What is a decision tree ?



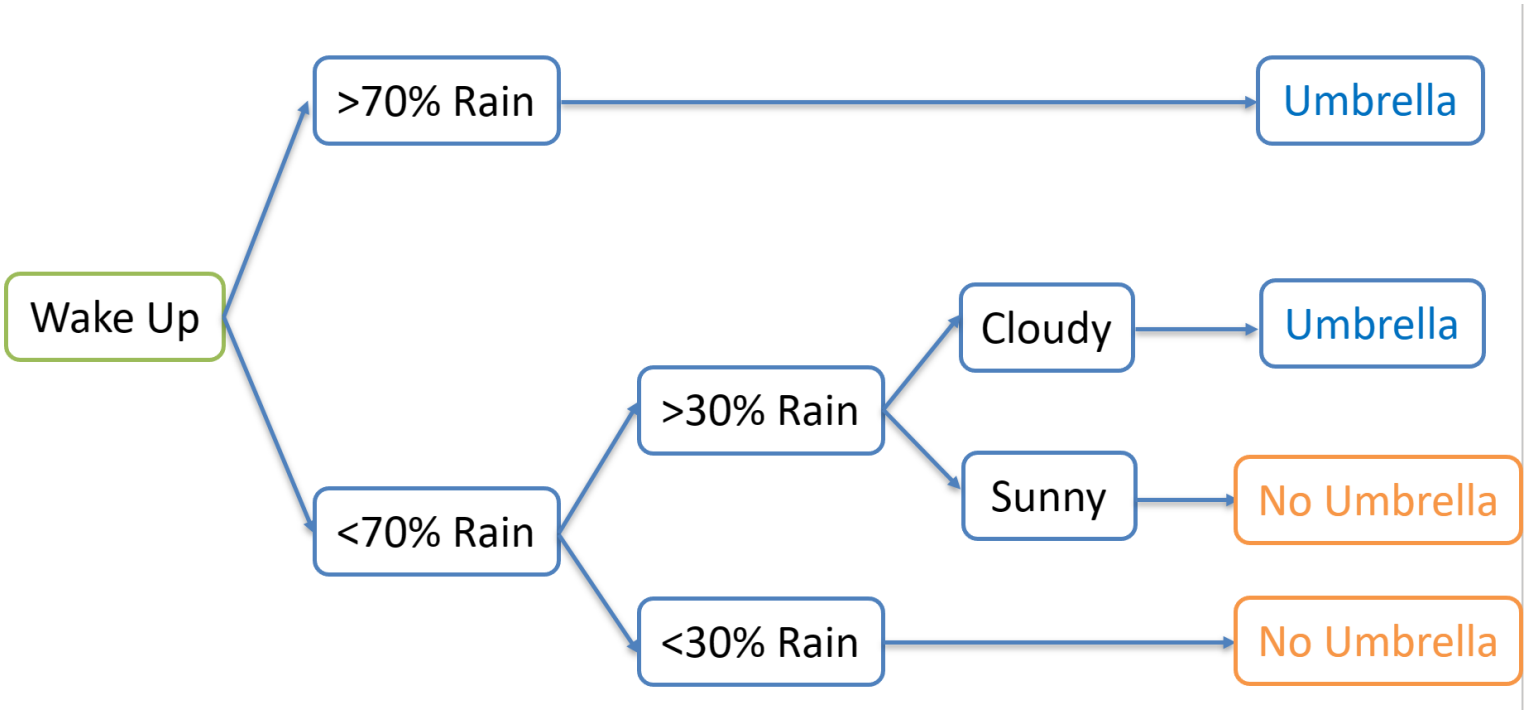
- Tree-logic uses a sequence of inquiries to come to a conclusion.

What is a decision tree ?



- Tree-logic uses a sequence of inquiries to come to a conclusion.
- The trick is to have mini-decisions combine for good choices.

What is a decision tree ?



- Tree-logic uses a sequence of inquiries to come to a conclusion.
- The trick is to have mini-decisions combine for good choices.
- Each decision is a node, and the final prediction is a leaf node.

Decision tree

- A graphical method to describe decision problems

Decision tree

- A graphical method to describe decision problems
- Illustrate key decision variables, time sequence, and available information:

Decision tree

- A graphical method to describe decision problems
- Illustrate key decision variables, time sequence, and available information:
  - Alternatives

Decision tree

- A graphical method to describe decision problems
- Illustrate key decision variables, time sequence, and available information:
  - Alternatives
  - Possible outcomes and associated probabilities for each alternative

Decision tree

- A graphical method to describe decision problems
- Illustrate key decision variables, time sequence, and available information:
  - Alternatives
  - Possible outcomes and associated probabilities for each alternative
  - Gains and losses of each possible outcome

Structure decision problems with a "Tree"

Decision trees have three types of nodes:

- A decision node (a square)

Structure decision problems with a "Tree"

Decision trees have three types of nodes:

- A decision node (a square)
  - When the decision maker makes a choice.

Structure decision problems with a "Tree"

Decision trees have three types of nodes:

- A decision node (a square)
  - When the decision maker makes a choice.
- A probability node (a circle), also called "chance" or "event" node

Structure decision problems with a "Tree"

Decision trees have three types of nodes:

- A decision node (a square)
  - When the decision maker makes a choice.
- A probability node (a circle), also called "chance" or "event" node
  - When the result of an uncertain event becomes known.

Structure decision problems with a "Tree"

Decision trees have three types of nodes:

- A decision node (a square)
  - When the decision maker makes a choice.
- A probability node (a circle), also called "chance" or "event" node
  - When the result of an uncertain event becomes known.
- An end node (a triangle), also called "leaf" node

Structure decision problems with a "Tree"

Decision trees have three types of nodes:

- A decision node (a square)
  - When the decision maker makes a choice.
- A probability node (a circle), also called "chance" or "event" node
  - When the result of an uncertain event becomes known.
- An end node (a triangle), also called "leaf" node
  - When the problem is completed (all decisions have been made, all uncertainty has been resolved, and all payoffs/costs have been incurred).

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node
  - Represent different alternatives open to the decision maker

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node
  - Represent different alternatives open to the decision maker
  - Decision maker chooses one alternative

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node
  - Represent different alternatives open to the decision maker
  - Decision maker chooses one alternative
- Probability branches:

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node
  - Represent different alternatives open to the decision maker
  - Decision maker chooses one alternative
- Probability branches:
  - Begin from a chance node

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node
  - Represent different alternatives open to the decision maker
  - Decision maker chooses one alternative
- Probability branches:
  - Begin from a chance node
  - Represent possible outcomes of an uncertain event

Structure decision problems with a "Tree"

Decision trees have two types of branches:

- Alternative branches:
  - Begin from a decision node
  - Represent different alternatives open to the decision maker
  - Decision maker chooses one alternative
- Probability branches:
  - Begin from a chance node
  - Represent possible outcomes of an uncertain event
  - Decision maker has no control over which will occur, but can specify a probability for each branch

Structure decision problems with a "Tree"

Time proceeds from left to right:

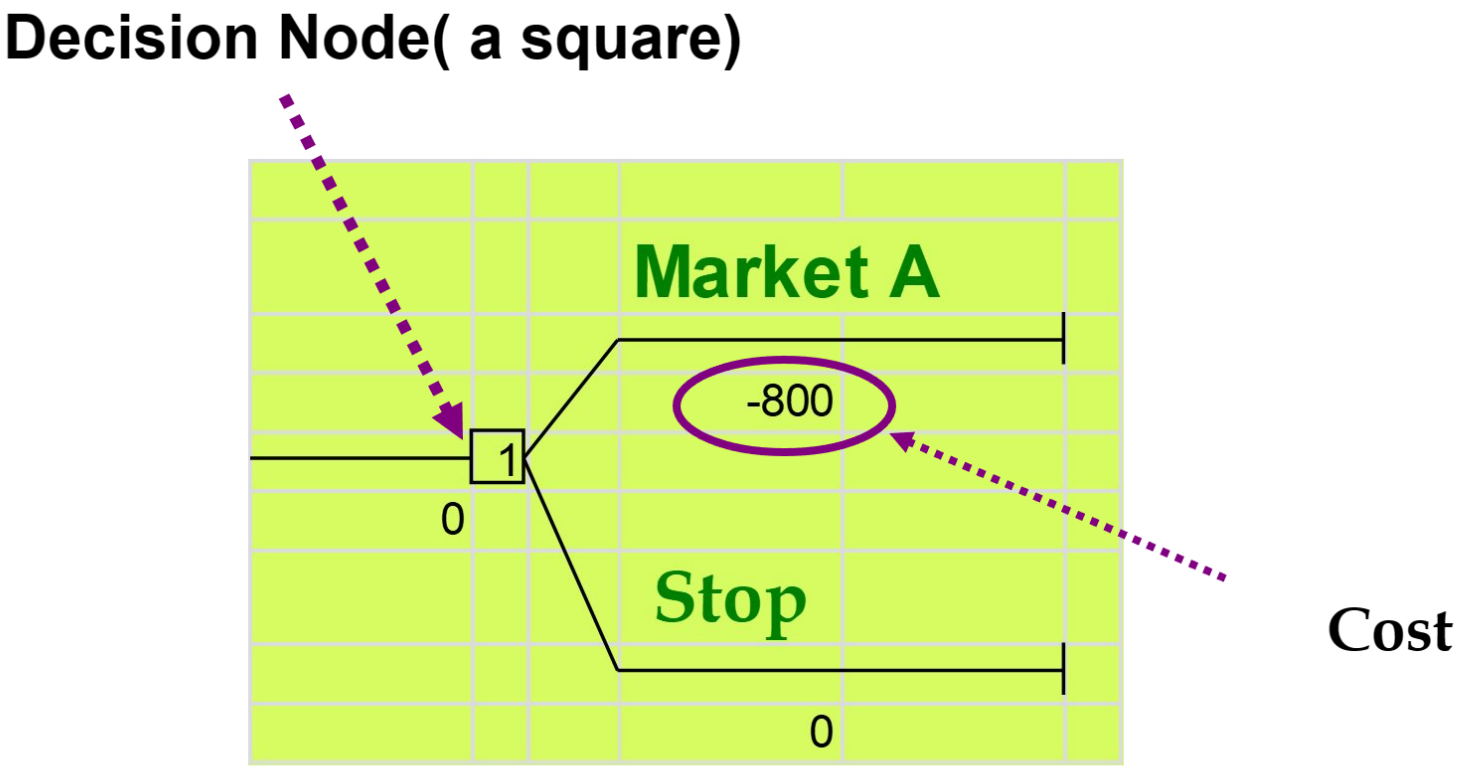
- Any branches leading into a node (from the left) have already occurred.

Structure decision problems with a "Tree"

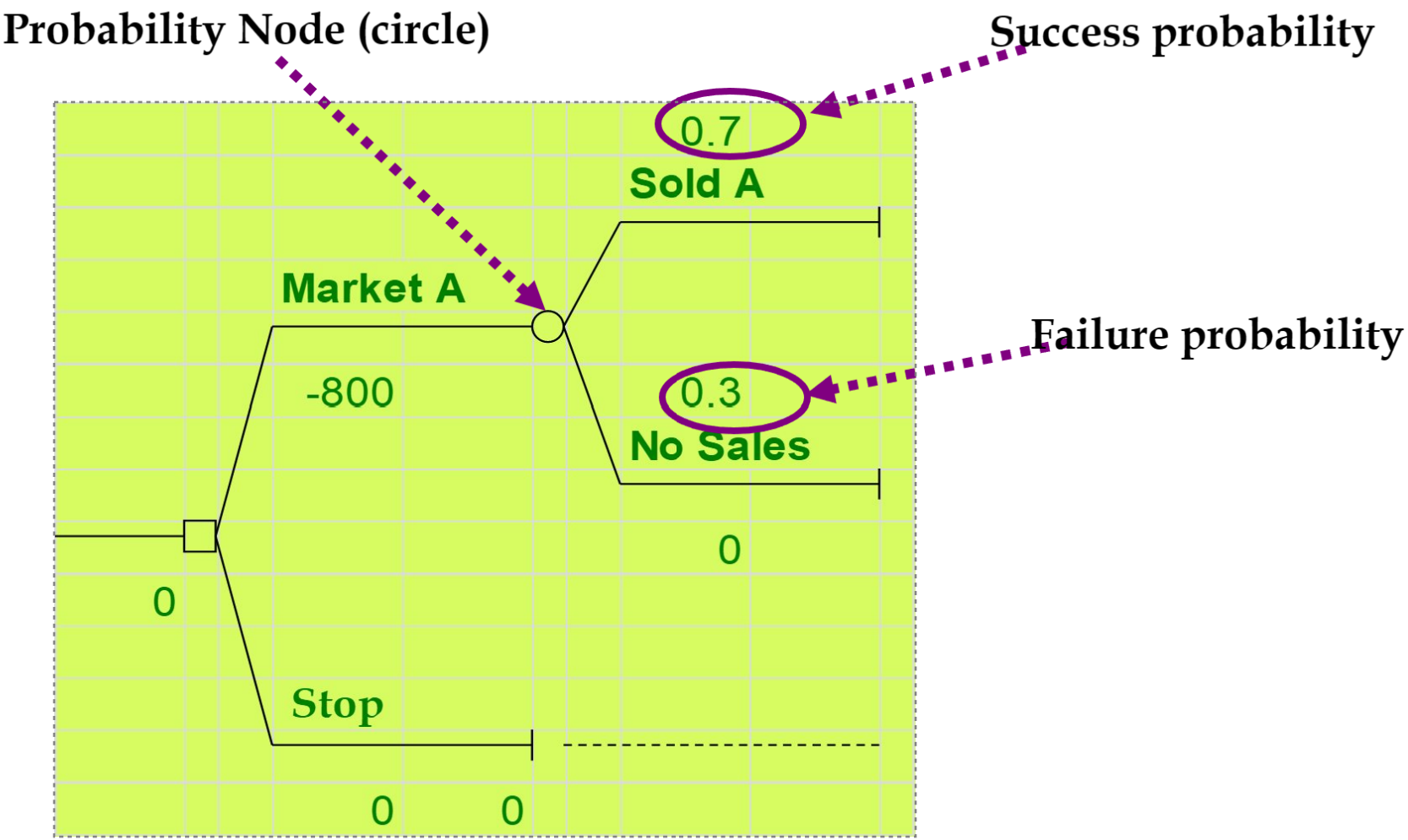
Time proceeds from left to right:

- Any branches leading into a node (from the left) have already occurred.
- Any branches leading out of a node (to the right) have not yet occurred.

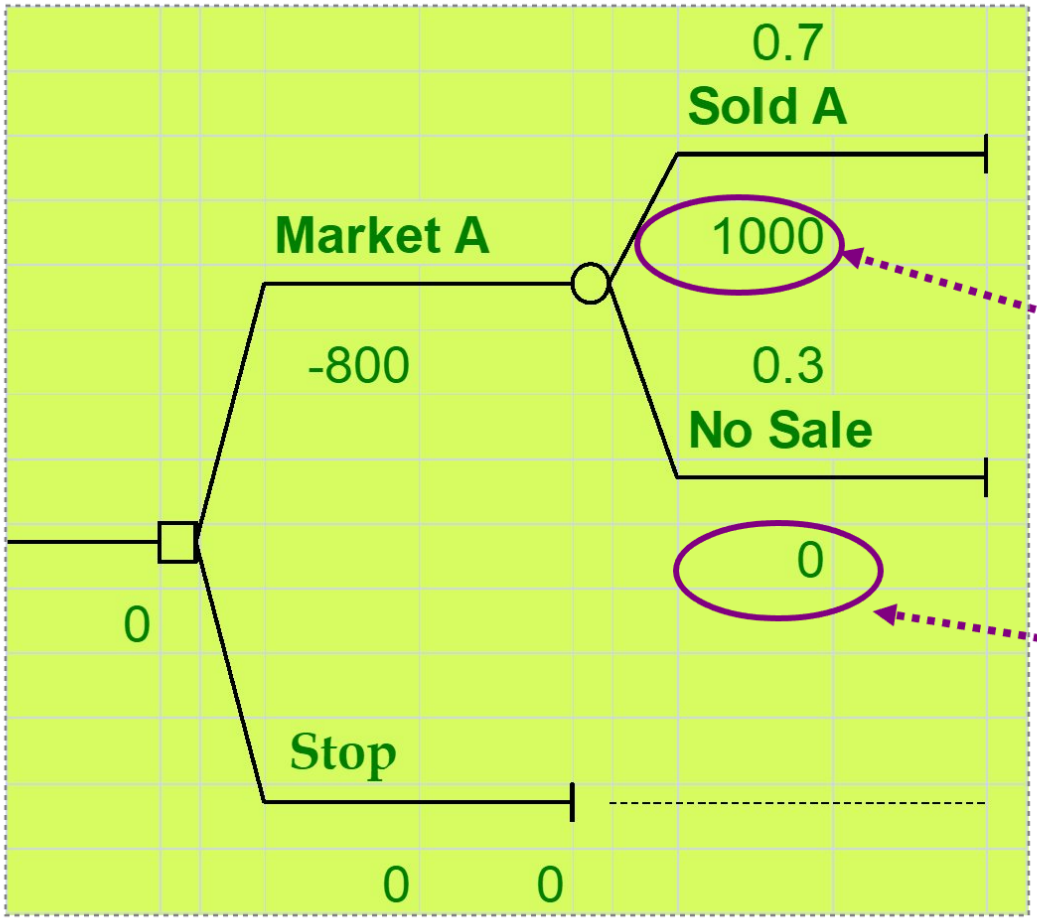
Construct a decision tree (decision node, cost)



Construct a decision tree (probability node, probability)



Construct a decision tree (all relative revenue and loss)



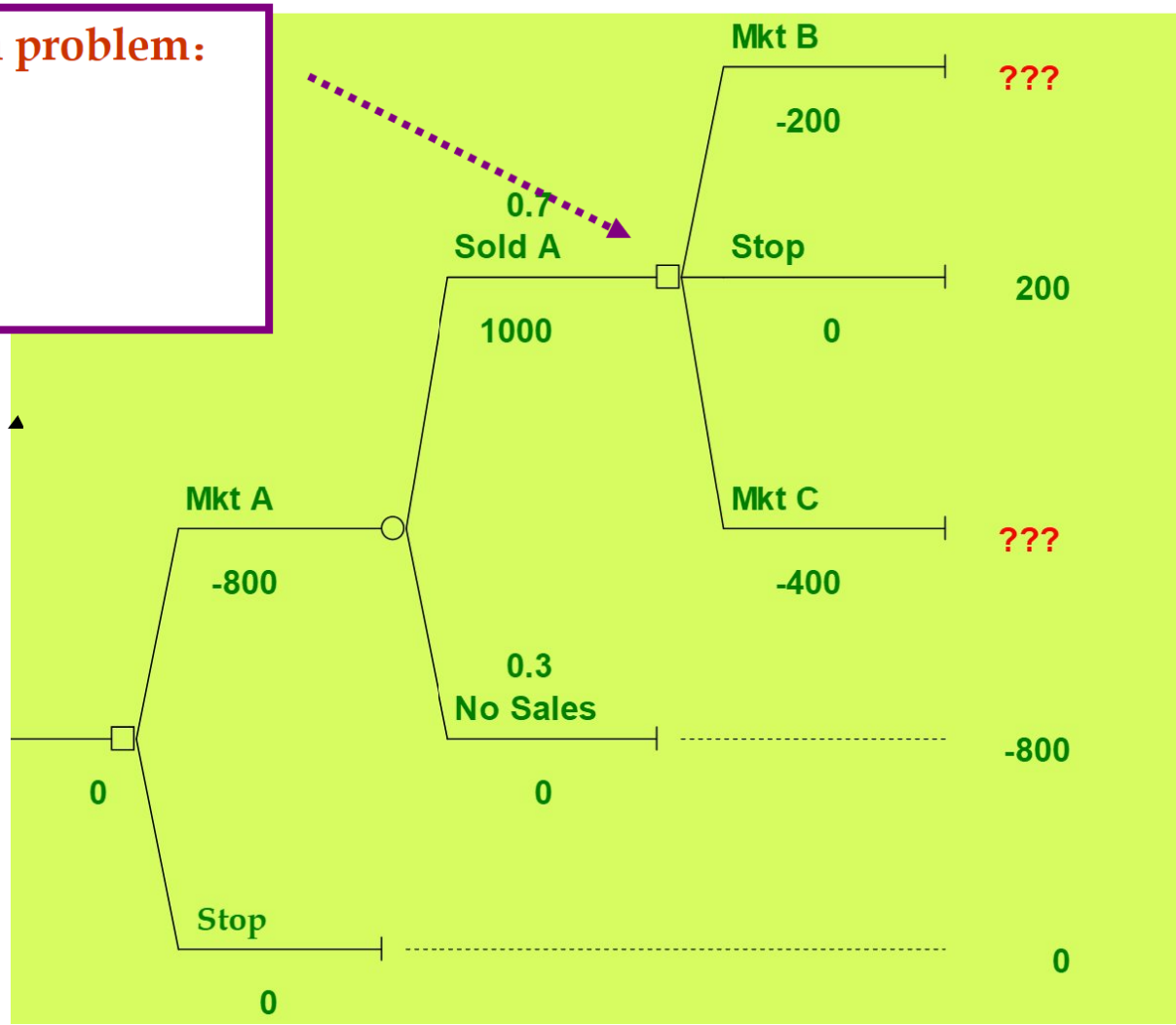
If sold, revenue
=(25,000)x(4%)
=1000

If no sales, no revenue & no right to sell other properties

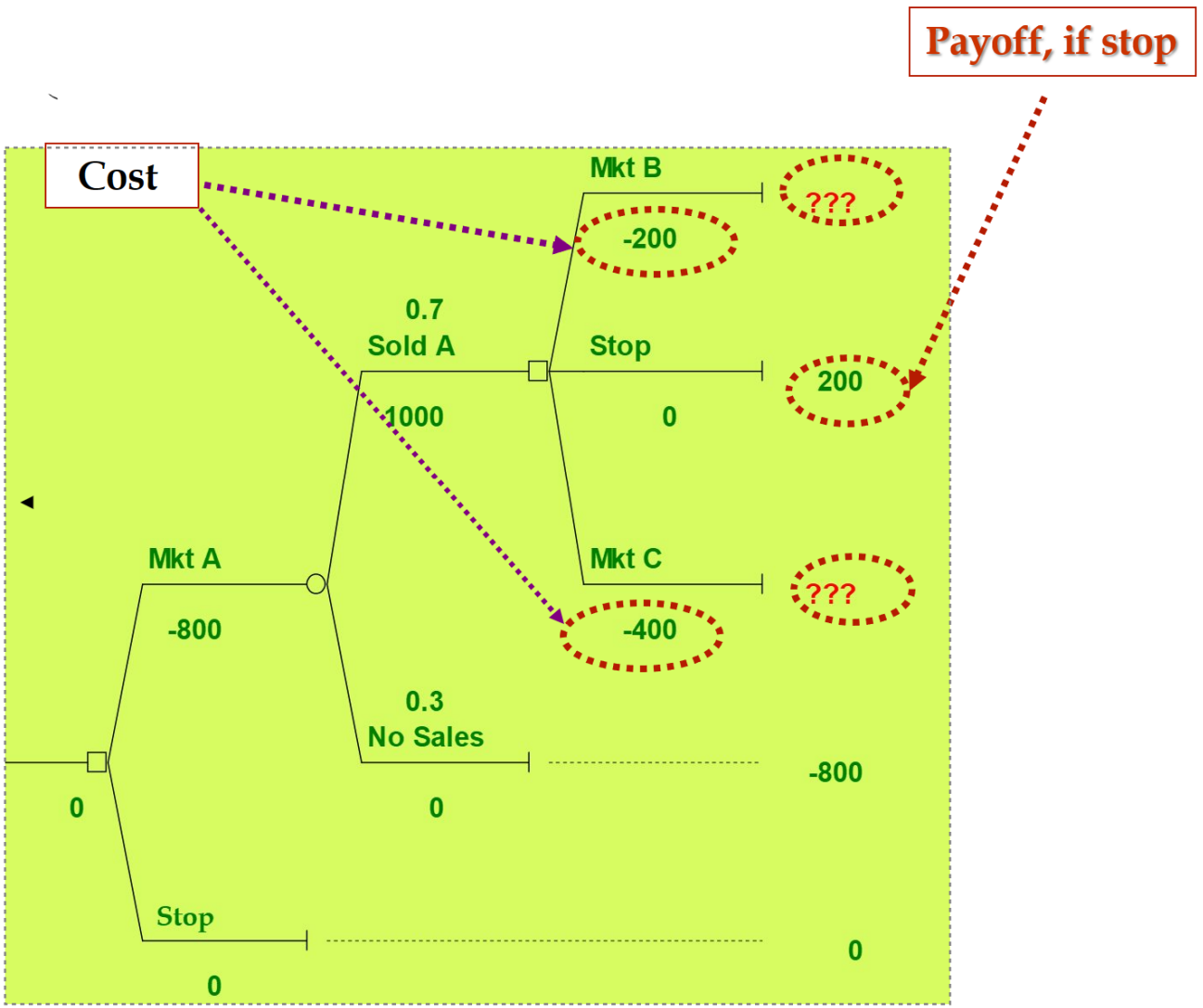
Construct a decision tree (multiple phase decisions)

Multiple Decision problem:
After A is sold →

- **Market B?**
- **Stop?**
- **Market C?**



Construct a decision tree (multiple phase decisions)



Group discussion 1

Grow a decision tree to describe

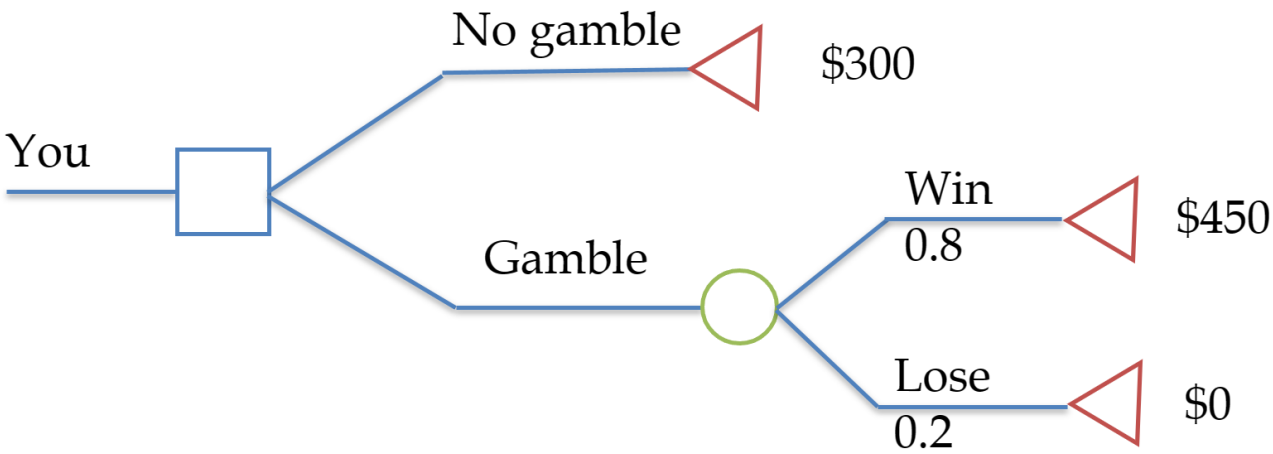
- win \$300 for sure

Group discussion 1

Grow a decision tree to describe

- win \$300 for sure
- play a gamble in which you have an 80% chance of winning \$450, otherwise \$0.

Group discussion 1



Group discussion 2

Grow a decision tree to describe

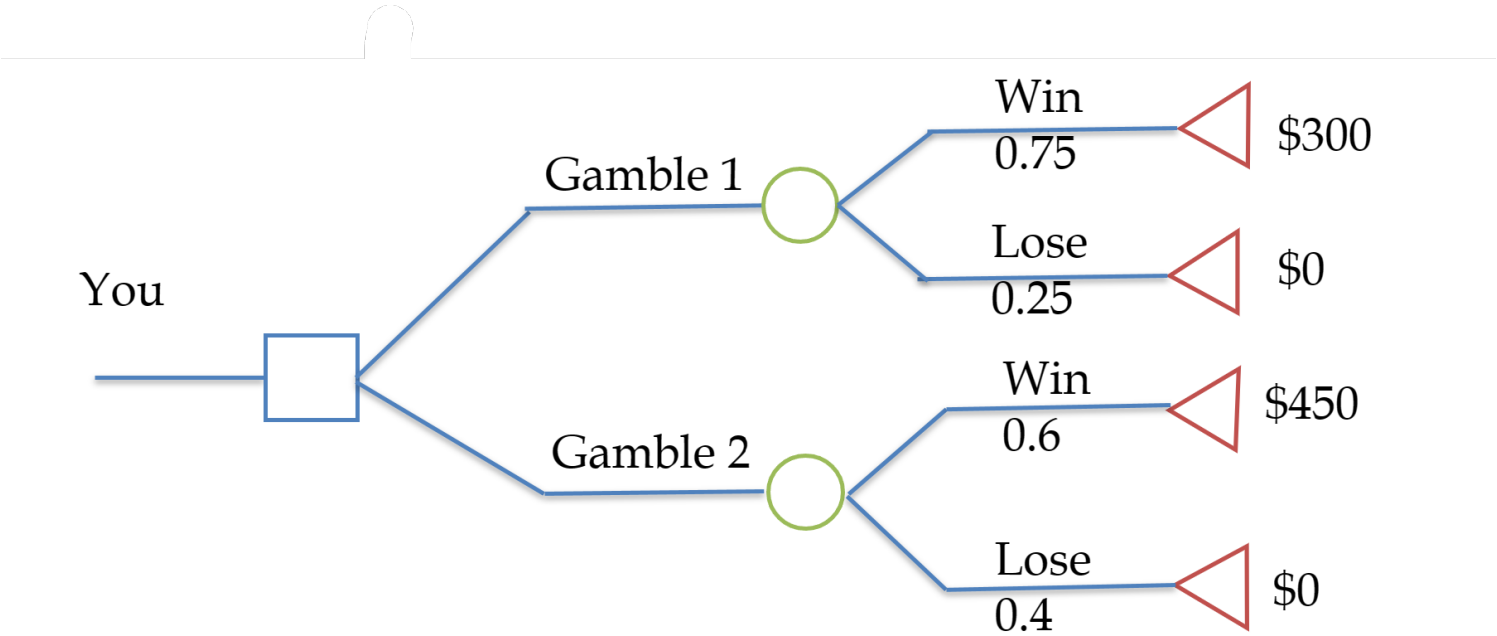
- play a gamble with a 75% chance of winning \$300, otherwise \$0

Group discussion 2

Grow a decision tree to describe

- play a gamble with a 75% chance of winning \$300, otherwise \$0
- play a gamble with a 60% chance of winning \$450, otherwise \$0.

Group discussion 2



Warren's decision tree

Selling Rule:

- Sell A first.
- If A is unsold within a month, the entire deal is off—no commission and no chance to sell the other properties.
- If A is sold within a month, then Warren will get the commission for A and the option of stopping at this point or trying to sell either the B or C next under the same conditions (i.e., sell within a month or no commission on the second property and no chance to sell the third).

	Cost	Price	Commission (4%)	Probability
A	800	25000	1000	0.7
B	200	50000	2000	0.6
C	400	100000	4000	0.5

Warren's decision tree

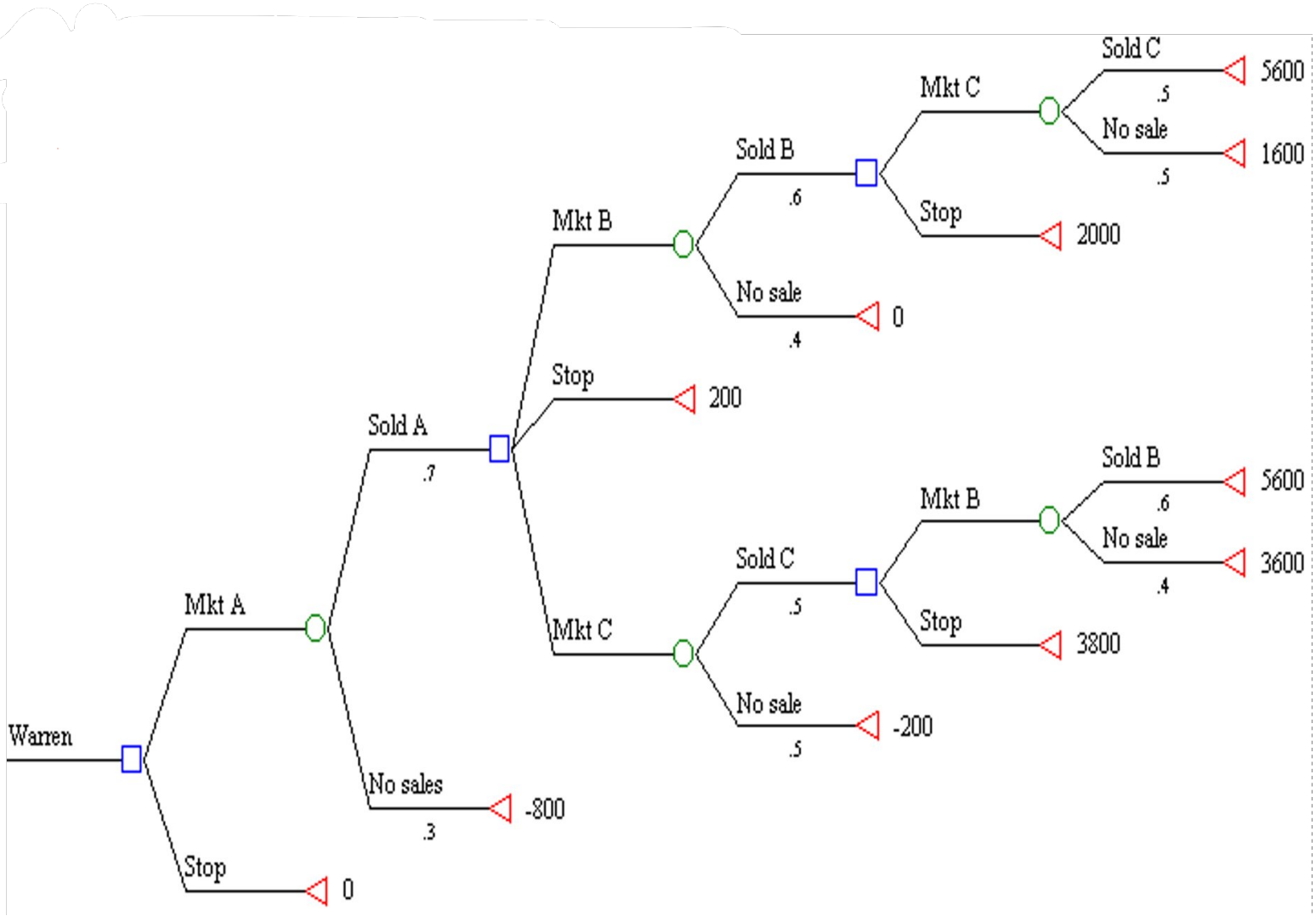
Selling Rule:

- Sell A first.
- If A is unsold within a month, the entire deal is off—no commission and no chance to sell the other properties.
- If A is sold within a month, then Warren will get the commission for A and the option of stopping at this point or trying to sell either the B or C next under the same conditions (i.e., sell within a month or no commission on the second property and no chance to sell the third).

	Cost	Price	Commission (4%)	Probability
A	800	25000	1000	0.7
B	200	50000	2000	0.6
C	400	100000	4000	0.5

Please try to construct Warren's tree with your classmates

Warren's decision tree



How to choose among multiple alternatives?

- Calculate Expected Monetary Value (EMV)!

Expected Monetary Value (EMV)

- Example: Investment

Expected Monetary Value (EMV)

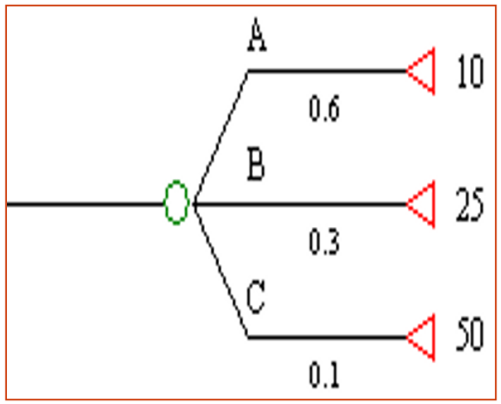
- Example: Investment
- 3 possible outcomes: A (bad), B(medium), C(good)

Expected Monetary Value (EMV)

- Example: Investment
- 3 possible outcomes: A (bad), B(medium), C(good)

Expected Monetary Value (EMV)

- Example: Investment
- 3 possible outcomes: A (bad), B(medium), C(good)



EMV:

$$\text{EMV} = \Sigma V_i P_i$$
$$= (10)(0.6) + (25)(0.3) + (50)(0.1)$$
$$= 18.5$$

Expected monetary value (EMV)

- Expected Monetary Value (EMV)—the weighted average of the possible outcomes, where the weights are the probabilities of those outcomes. Formally, if V_i is the monetary value corresponding to outcome i and P_i is its probability, then expected monetary value is defined as

$$EMV = \sum V_i P_i.$$

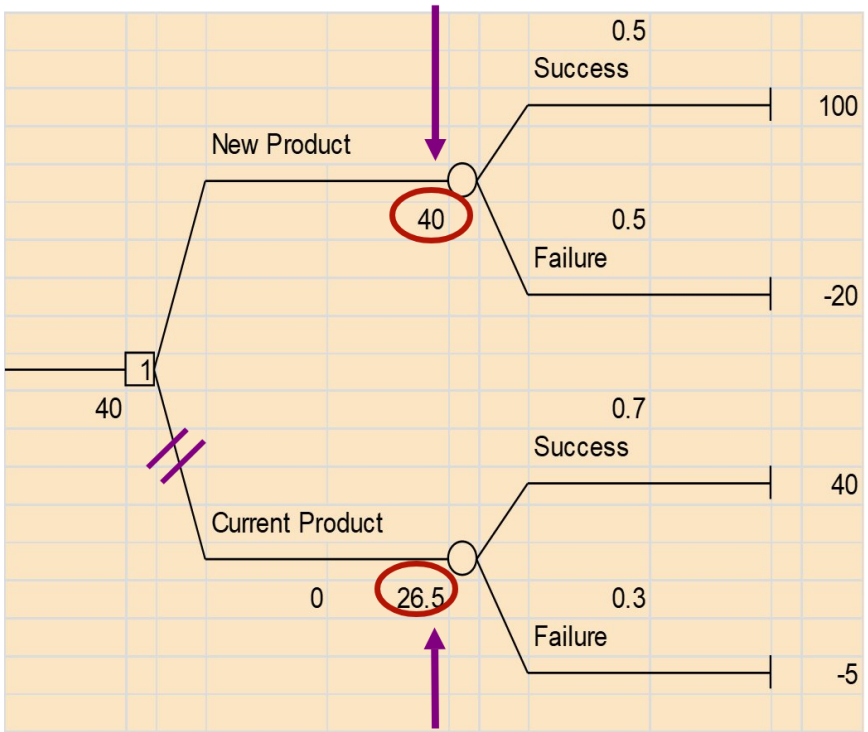
Expected monetary value (EMV)

- Expected Monetary Value (EMV)—the weighted average of the possible outcomes, where the weights are the probabilities of those outcomes. Formally, if V_i is the monetary value corresponding to outcome i and P_i is its probability, then expected monetary value is defined as
$$EMV = \sum V_i P_i.$$
- One common way to make the choice is to calculate EMV of each alternative and then choose the alternative with the largest EMV.

Making choice using EMV

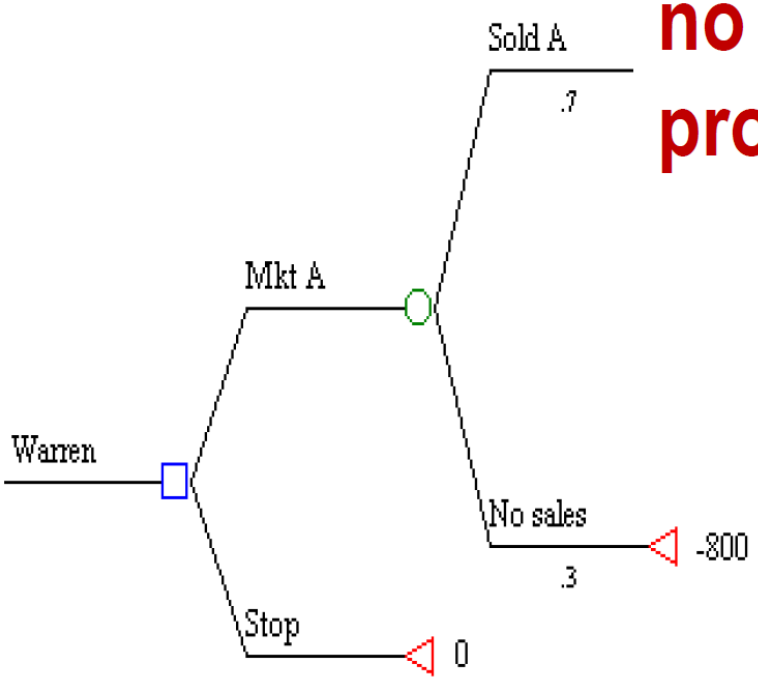
- Maximum EMV Criterion—using the information given in the payoffs together with the information given in the probabilities to determine the decision that has the highest expected value.

EMV=(100)*(0.5)+(-20)*(0.5)=40→optimal



EMV=(40)*(0.7)+(-5)*(0.3)=26.5

How to make a choice?



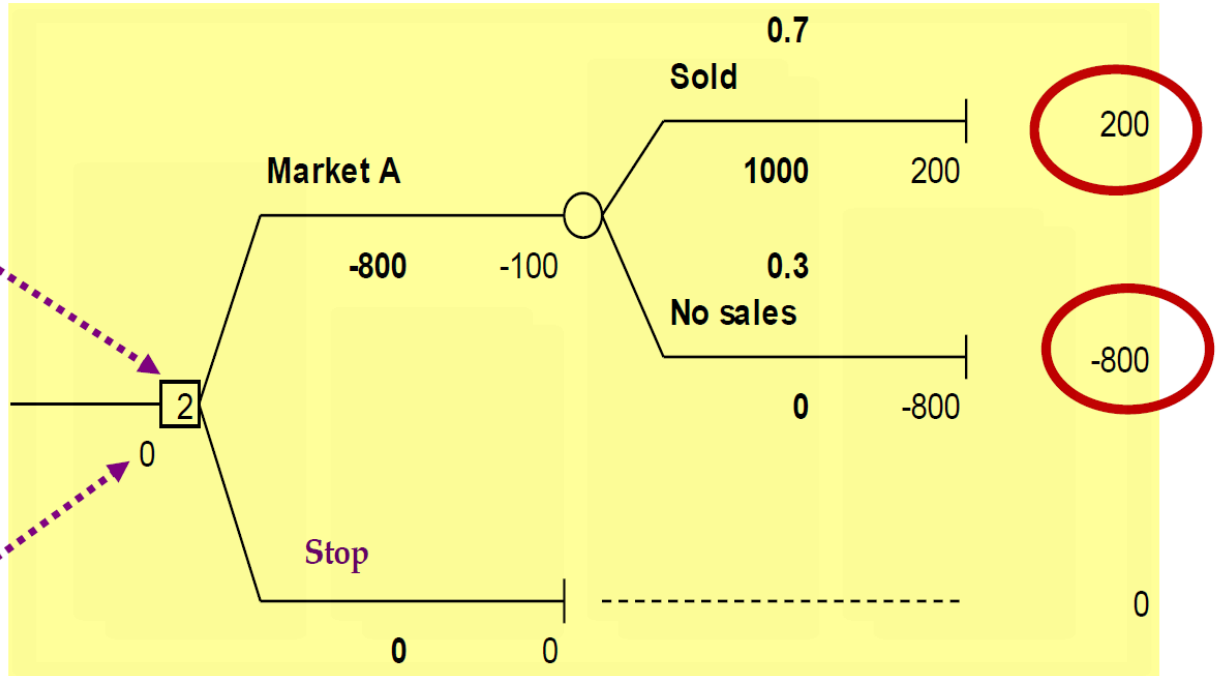
Let's first consider a simple single stage decision: no chance to sell other properties

Single stage decision

$$\text{EMV}=(25{,}000\times 4\%-800)(0.7)+(-800)(0.3)=\textbf{-100} \rightarrow \textbf{Loss!!}$$

The number in the decision node indicates which branch is optimal based on EMV.

The number before the decision node indicates EMV if following the optimal decision.



Group discussion 3

- Which to choose

Group discussion 3

- Which to choose
  - win $ 300 for sure

Group discussion 3

- Which to choose
  - win $ 300 for sure
  - play a gamble in which you have an 80% chance of winning $ 450, otherwise $0

Group discussion 3

- Which to choose
  - win $ 300 for sure
  - play a gamble in which you have an 80% chance of winning $ 450, otherwise $0
- Which to choose

Group discussion 3

- Which to choose
  - win $ 300 for sure
  - play a gamble in which you have an 80% chance of winning $ 450, otherwise $0
- Which to choose
  - play a gamble with a 75% chance of winning $300, otherwise $0

Group discussion 3

- Which to choose
  - win $ 300 for sure
  - play a gamble in which you have an 80% chance of winning $ 450, otherwise $0
- Which to choose
  - play a gamble with a 75% chance of winning $300, otherwise $0
  - play a gamble with a 60% chance of winning $450, otherwise $0

A multiple stage decision

- Warren's case $\rightarrow$ multiple stage problem:
  If A can be sold within a month $\rightarrow$ potential to earn profits!

A multiple stage decision

- Warren's case $\rightarrow$ multiple stage problem:
  If A can be sold within a month $\rightarrow$ potential to earn profits!
  - How big is the chance?

A multiple stage decision

- Warren's case $\rightarrow$ multiple stage problem:
  If A can be sold within a month $\rightarrow$ potential to earn profits!
  - How big is the chance?
  - Is the risk worth taking?

A multiple stage decision

- Warren's case $\rightarrow$ multiple stage problem:
  If A can be sold within a month $\rightarrow$ potential to earn profits!
  - How big is the chance?
  - Is the risk worth taking?
- "Folding back the tree" $\rightarrow$ Solve for the optimal strategy

A multiple stage decision

- Warren's case $\rightarrow$ multiple stage problem:
  If A can be sold within a month $\rightarrow$ potential to earn profits!
  - How big is the chance?
  - Is the risk worth taking?
- "Folding back the tree" $\rightarrow$ Solve for the optimal strategy
- Decision software (such as TreePlan in Excel) automatically computes all decision branches' EMV, and identifies the optimal choice.

Solving a decision tree: folding back

- Start at the far right side of tree and move left

Solving a decision tree: folding back

- Start at the far right side of tree and move left
- At chance nodes: calculate EMV

Solving a decision tree: folding back

- Start at the far right side of tree and move left
- At chance nodes: calculate EMV
- At decision modes:

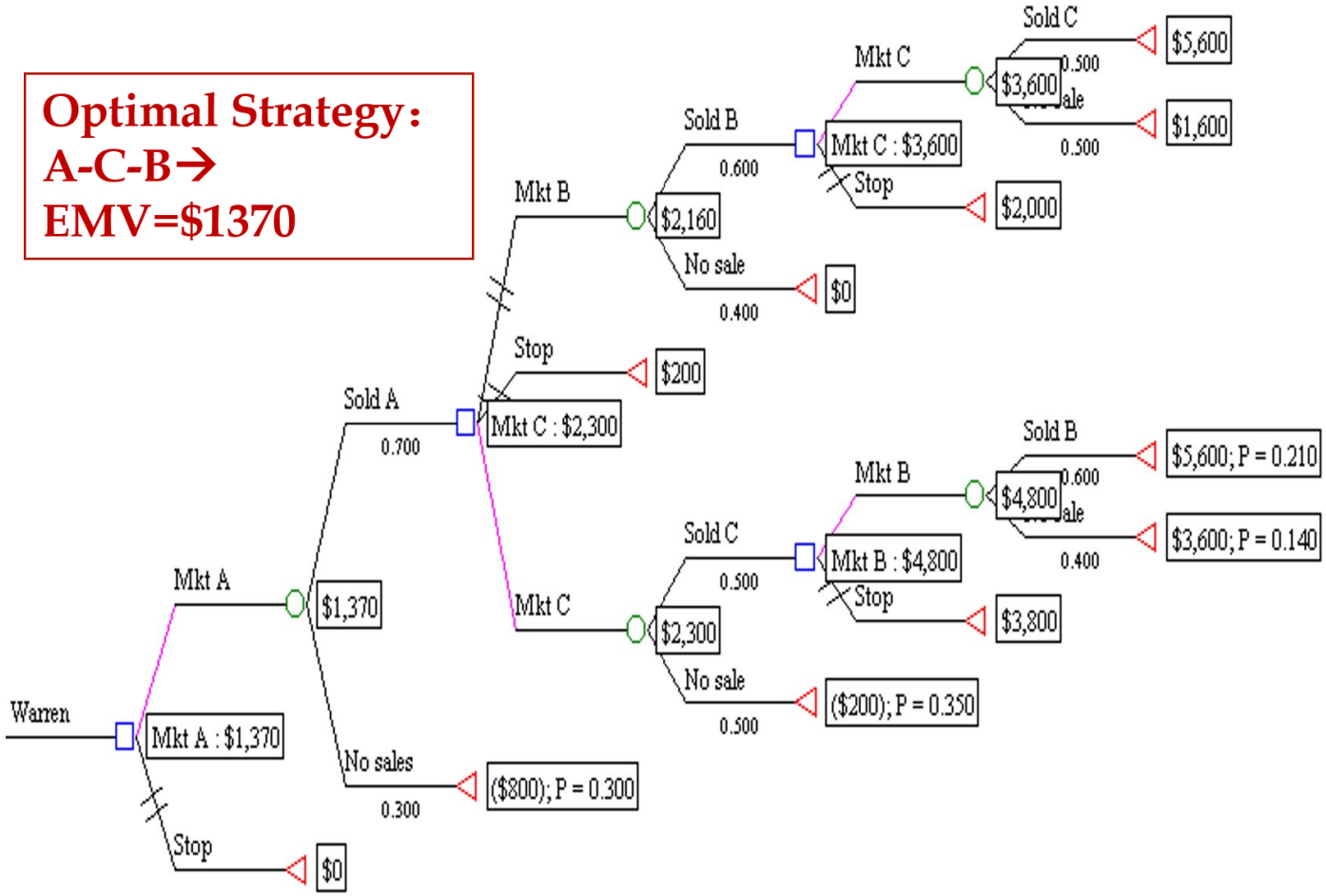
Solving a decision tree: folding back

- Start at the far right side of tree and move left
- At chance nodes: calculate EMV
- At decision modes:
  - Compare EMV of the alternatives

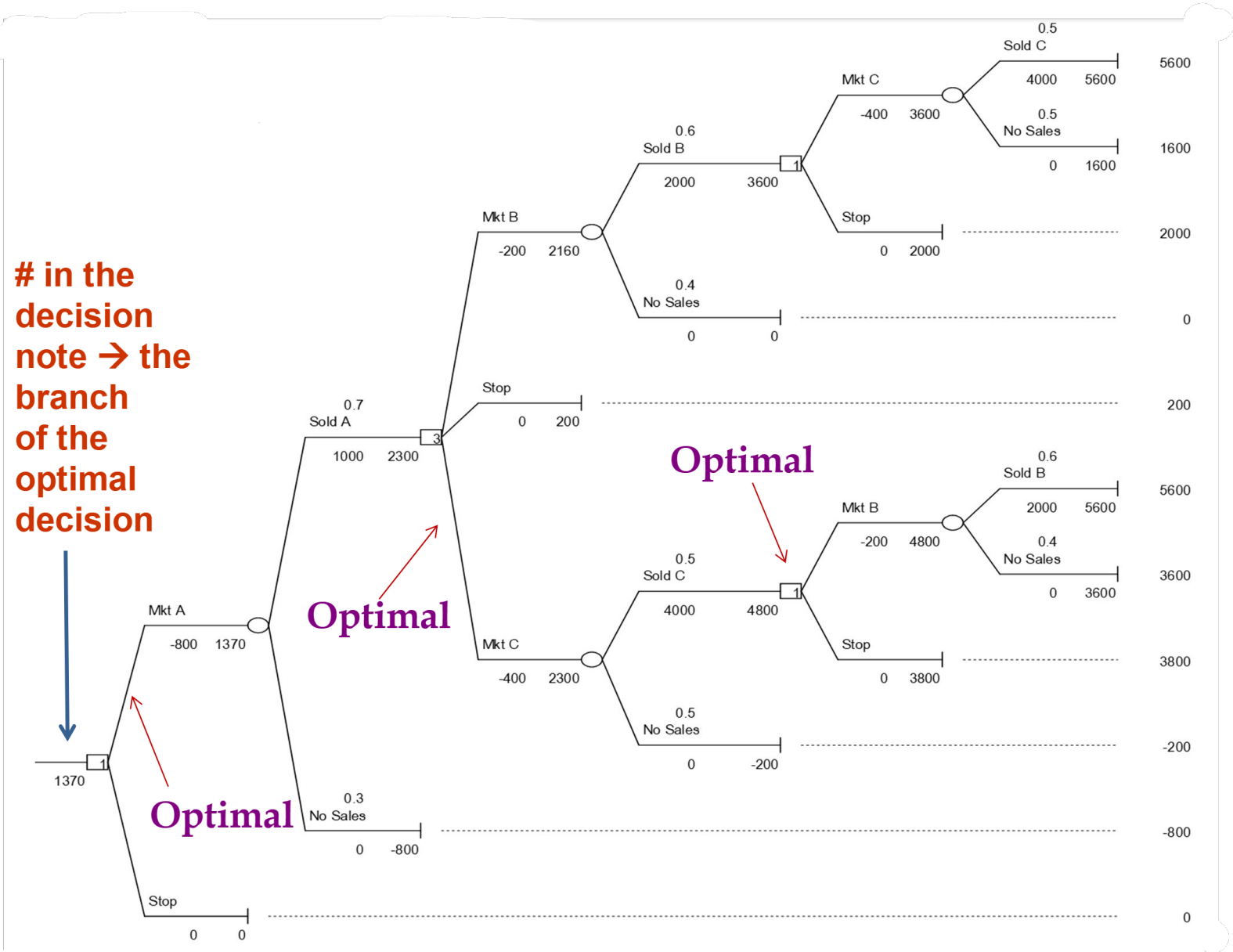
Solving a decision tree: folding back

- Start at the far right side of tree and move left
- At chance nodes: calculate EMV
- At decision modes:
  - Compare EMV of the alternatives
  - Select the alternative with the highest EMV

"Folding back the tree"



Building A decision tree using TreePlan



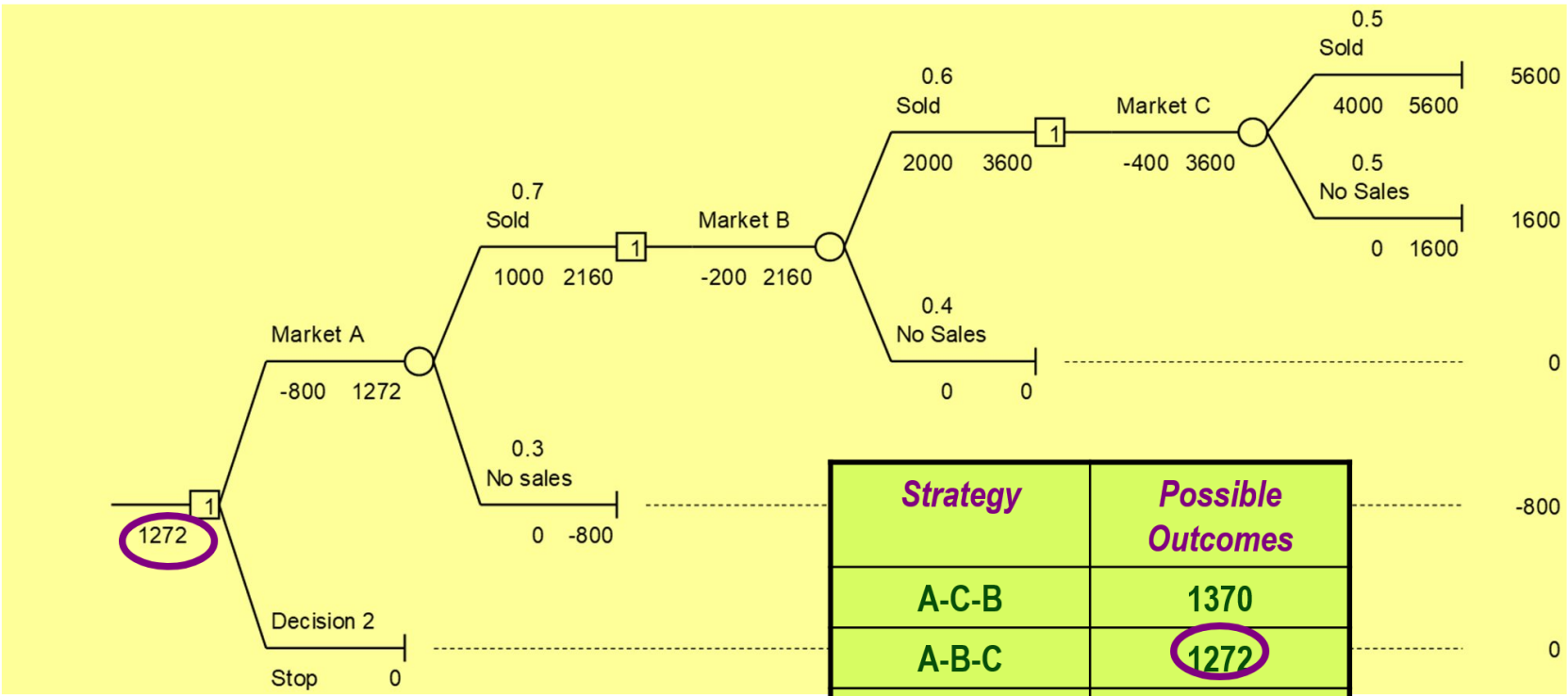
Warren's decisions

- Compare EMV of different alternative strategies

Strategy	*EMV*
A-C-B	**1370**
A-B-C	**1272**
A-C-Stop	**1020**
A-B-Stop	**600**
A-Stop	**-100**

Group discussion: EMV for other alternatives?

- please verify "A-B-C" strategy: EMV = $ 1272



Strategy	Possible Outcomes
A-C-B	1370
A-B-C	1272
A-C-Stop	1020
A-B-Stop	600
A-Stop	-100

Case: new product introduction

Product A

- A problem with the production system has not yet been solved: P(delay) $= 0.05$

Case: new product introduction

Product A

- A problem with the production system has not yet been solved: P(delay) $= 0.05$
- Price: High or Low (The price would not be set until just before the product is to be introduced.)

Case: new product introduction

Product A

- A problem with the production system has not yet been solved: P(delay) $= 0.05$
- Price: High or Low (The price would not be set until just before the product is to be introduced.)

Case: new product introduction

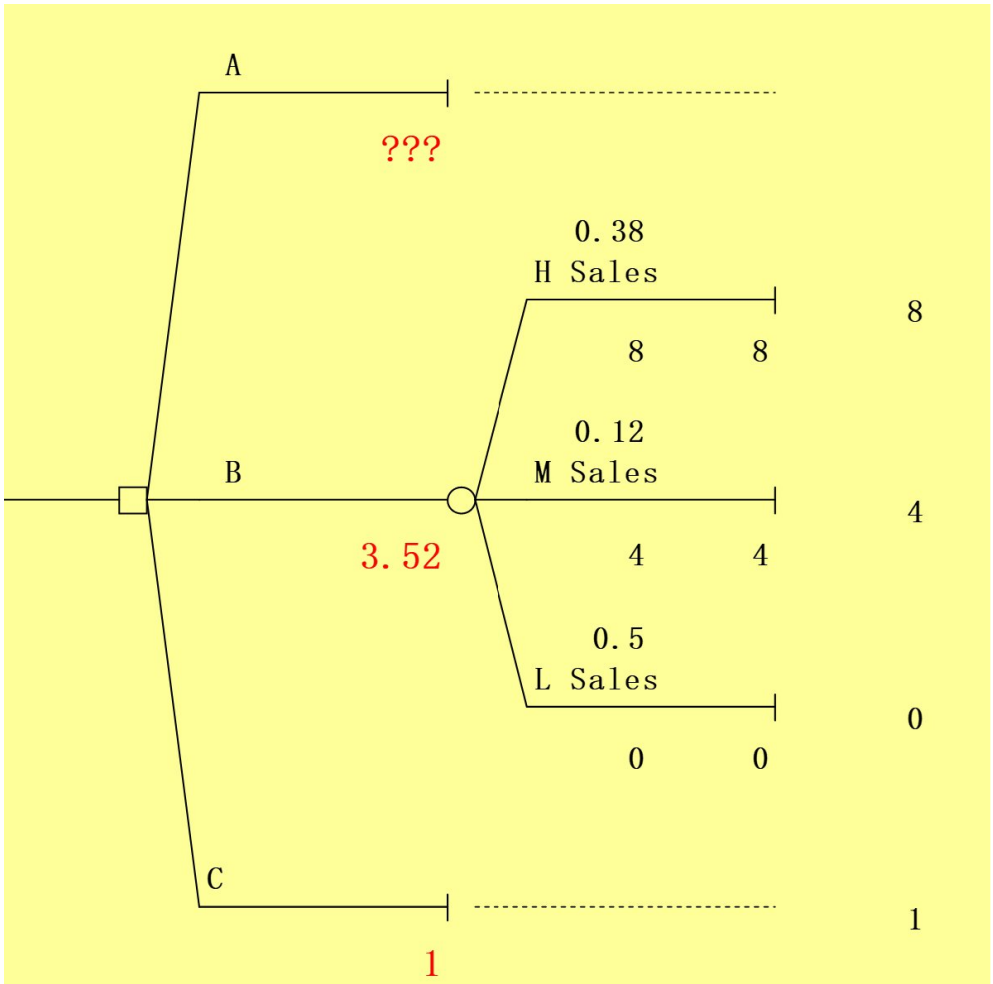
Product A

- A problem with the production system has not yet been solved: P(delay) = 0.05
- Price: High or Low (The price would not be set until just before the product is to be introduced.)

	Price	Payoff if High Sales (chance)	Payoff if Low Sales (chance)
Delay	High	5.0 (30%)	-0.5 (70%)
	Low	3.5 (50%)	1.0 (50%)
No Delay	High	12 (40%)	0.0 (60%)
	Low	4.5 (50%)	1.5 (50%)

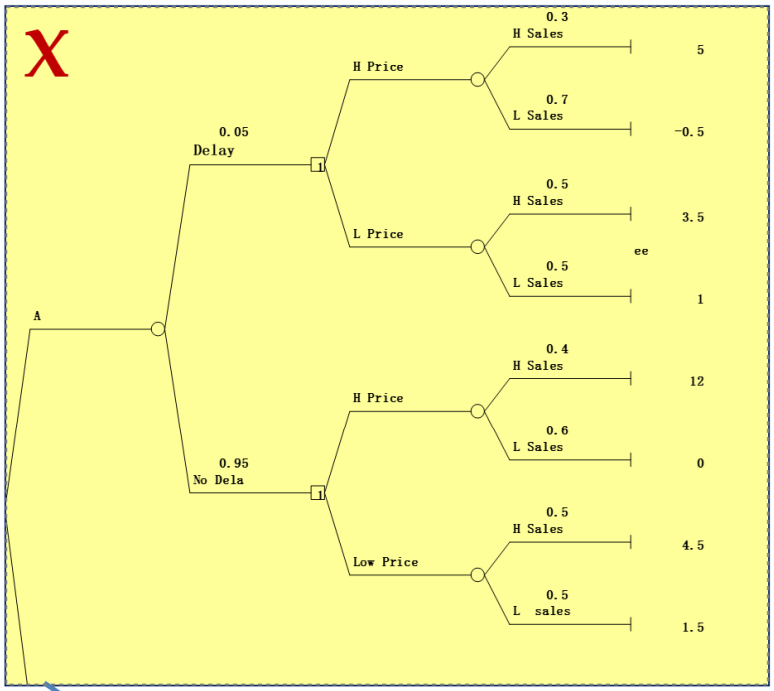
Case: new product introduction

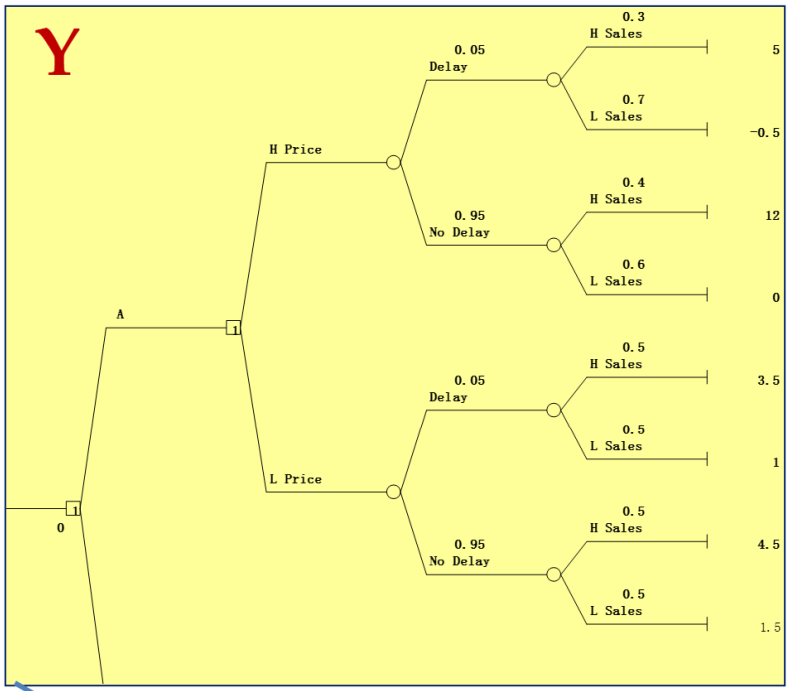
- Which product should be introduced?



Decision tree for product A

- Are the two trees the same?





Make decision with information

Make decision without information

Which tree is correct?

- Key Question: Do you have the information at the time when the decision has to be made? (Do you know if the process is delayed when you make the price decision)?

Which tree is correct?

- Key Question: Do you have the information at the time when the decision has to be made? (Do you know if the process is delayed when you make the price decision)?
  - If you have such information $\rightarrow X$ is correct (the decision node follows the chance node)

Which tree is correct?

- Key Question: Do you have the information at the time when the decision has to be made? (Do you know if the process is delayed when you make the price decision)?
  - If you have such information $\rightarrow X$ is correct (the decision node follows the chance node)
  - If you do not have such information $\rightarrow Y$ is correct (chance node follows the decision node).

Case: the risk of flooding

- The risk of flooding in land next to the River has recently increased.

Case: the risk of flooding

- The risk of flooding in land next to the River has recently increased.
- A tidal barrier is being constructed at the mouth of the river, but the Hartland River Authority has to decide how to provide flood protection in the two years before the barrier is ready.

Case: the risk of flooding

- The risk of flooding in land next to the River has recently increased.
- A tidal barrier is being constructed at the mouth of the river, but the Hartland River Authority has to decide how to provide flood protection in the two years before the barrier is ready.
- If flooding occurs in any one year, the Authority will have to pay out compensation of about $2 million for that year.

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.
  - 37% chance that the river's height will exceed 9.5 feet in any year.

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.
  - 37% chance that the river's height will exceed 9.5 feet in any year.
- Erect a cheap temporary barrier of 11 feet (first year only).

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.
  - 37% chance that the river's height will exceed 9.5 feet in any year.
- Erect a cheap temporary barrier of 11 feet (first year only).
  - Cost: $0.9 million

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.
  - 37% chance that the river's height will exceed 9.5 feet in any year.
- Erect a cheap temporary barrier of 11 feet (first year only).
  - Cost: \$0.9 million
  - 9% chance that the river's height will exceed 11 feet in any year

Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.
  - 37% chance that the river's height will exceed 9.5 feet in any year.
- Erect a cheap temporary barrier of 11 feet (first year only).
  - Cost: $0.9 million
  - 9% chance that the river's height will exceed 11 feet in any year
  - 30% chance of serious barrier damage if water tops barrier in year one, rendering it totally ineffective for year two

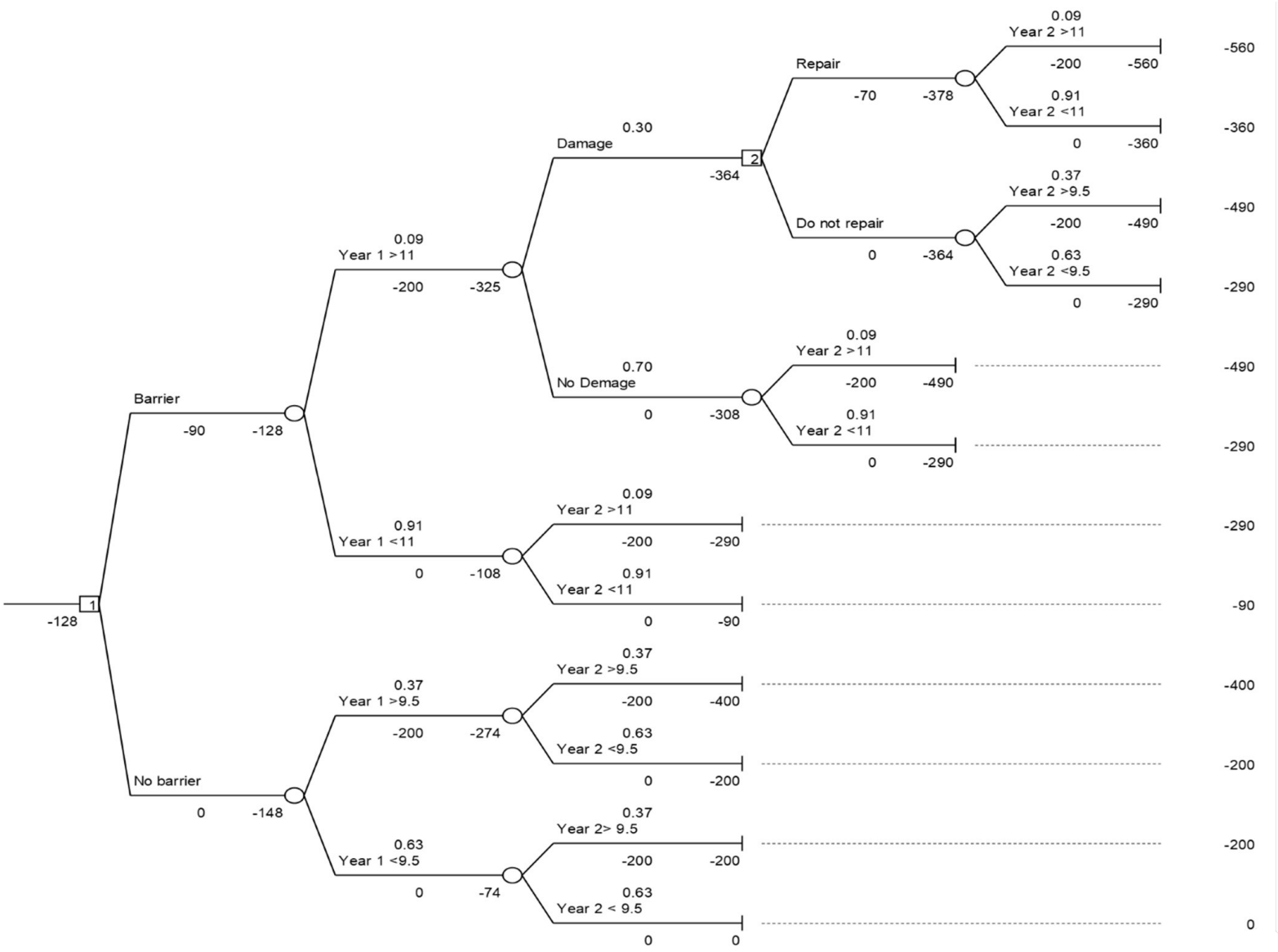
Case: the risk of flooding

The Hartland River Authority is considering two options:

- Do nothing
  - The river's natural banks will stop flooding as long as the height of the water is less than 9.5 feet.
  - 37% chance that the river's height will exceed 9.5 feet in any year.
- Erect a cheap temporary barrier of 11 feet (first year only).
  - Cost: \$0.9 million
  - 9% chance that the river's height will exceed 11 feet in any year
  - 30% chance of serious barrier damage if water tops barrier in year one, rendering it totally ineffective for year two
  - If damage, one can repair the barrier at a cost of \$0.7 million or leave the river unprotected for the second year

Case: the risk of flooding

after class exercise!



How to measure the effectiveness of a classifier ?

- Use ROC and AUC !

What is ROC?

- ROC stands for Receiver Operating Characteristic.

What is ROC?

- ROC stands for Receiver Operating Characteristic.
- The ROC curve is a graphical representation of a classifier's performance with x-axis and y-axis as follows:

What is ROC?

- ROC stands for Receiver Operating Characteristic.
- The ROC curve is a graphical representation of a classifier's performance with x-axis and y-axis as follows:
  - Y-axis: True Positive Rate (TPR), defined as the ratio of true positive observations to the total actual positives. It is expressed as:

$$TPR = \frac{TP}{TP + FN}.$$

What is ROC?

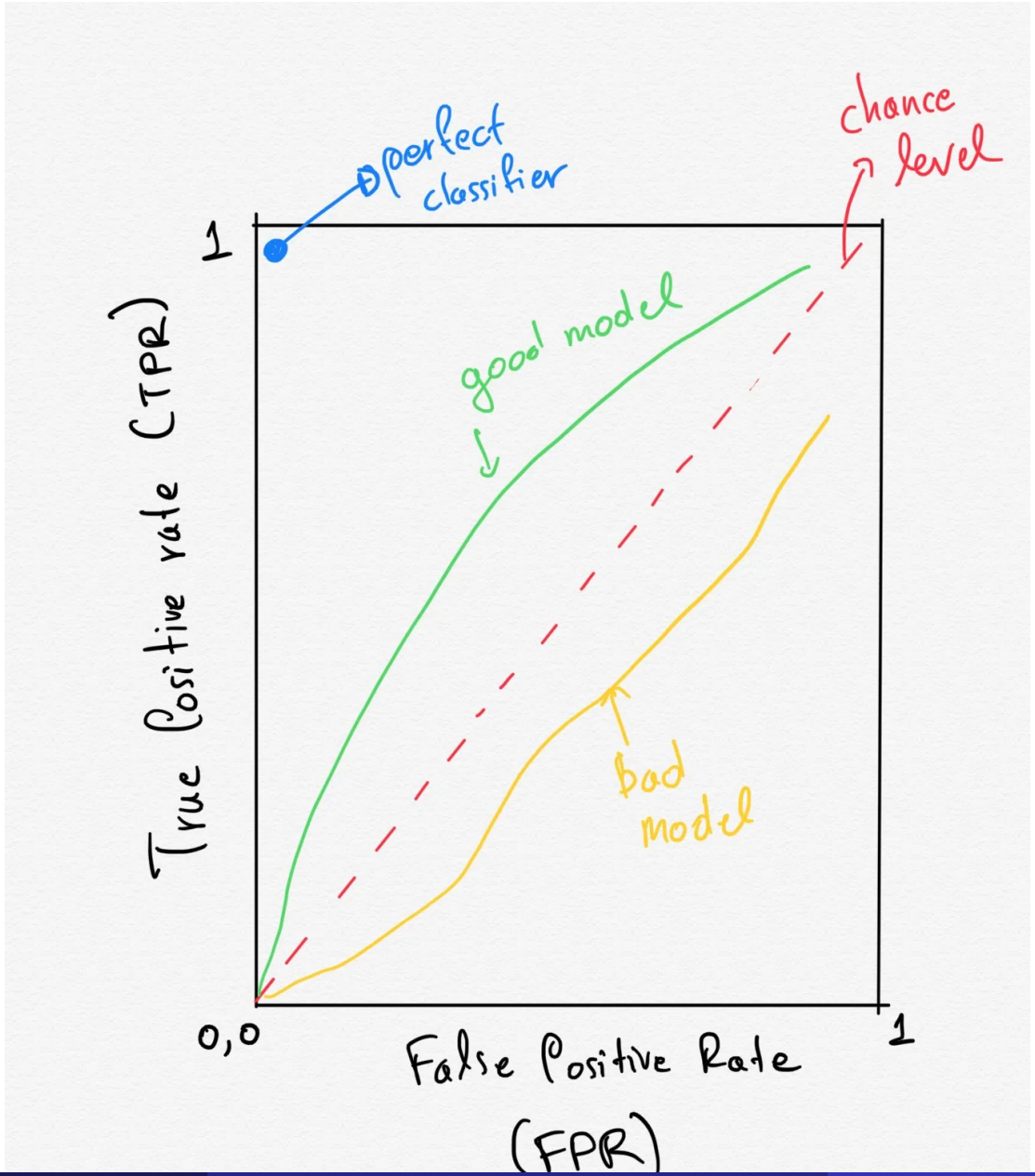
- ROC stands for Receiver Operating Characteristic.
- The ROC curve is a graphical representation of a classifier's performance with x-axis and y-axis as follows:
  - Y-axis: True Positive Rate (TPR), defined as the ratio of true positive observations to the total actual positives. It is expressed as:

$$TPR = \frac{TP}{TP + FN}.$$

  - X-axis: False Positive Rate (FPR), defined as the ratio of false positive observations to the total actual negatives. It is expressed as:

$$FPR = \frac{FP}{FP + TN}.$$

An example of ROC curve



What is AUC?

- AUC means the area under the curve

What is AUC?

- AUC means the area under the curve
- AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$. It provides an aggregate measure of performance across all possible classification thresholds.

What is AUC?

- AUC means the area under the curve
- AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$. It provides an aggregate measure of performance across all possible classification thresholds.
- The AUC can be interpreted as follows:

What is AUC?

- AUC means the area under the curve
- AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$. It provides an aggregate measure of performance across all possible classification thresholds.
- The AUC can be interpreted as follows:
  - A model whose predictions are 100% wrong has an AUC of 0.0;

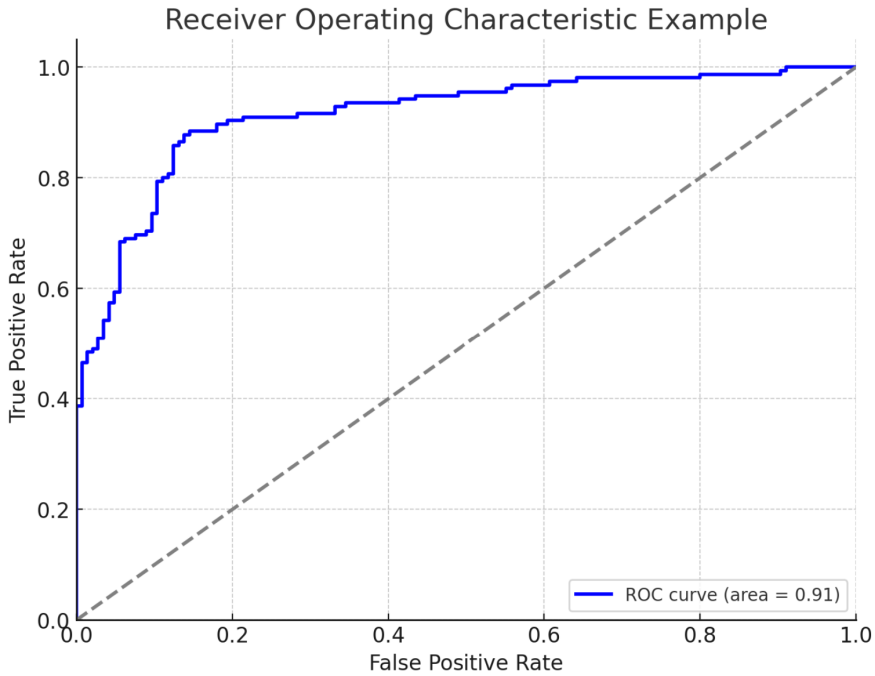
What is AUC?

- AUC means the area under the curve
- AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$. It provides an aggregate measure of performance across all possible classification thresholds.
- The AUC can be interpreted as follows:
  - A model whose predictions are 100% wrong has an AUC of 0.0;
  - A model whose predictions are 100% correct has an AUC of 1.0.

What is AUC?

- AUC means the area under the curve
- AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$. It provides an aggregate measure of performance across all possible classification thresholds.
- The AUC can be interpreted as follows:
  - A model whose predictions are 100% wrong has an AUC of 0.0;
  - A model whose predictions are 100% correct has an AUC of 1.0.
  - AUC values between 0 and 1 represent the degree of separability the model has. The higher the AUC, the better the model is at distinguishing between the positive and negative classes.

An example of ROC and AUC



- The blue line represents the ROC curve for a Logistic Regression model trained on a binary classification dataset. The area under the curve (AUC) is 0.91. The gray dashed line represents a baseline model that randomly guesses the class (AUC = 0.5). The closer the ROC curve follows the top-left border, the more accurate the model.

ROC and AUC in Decision Trees - Practical Example

- Suppose we're analyzing data from a medical study to predict the risk of a certain disease. We'll use one key health metric as our feature and the presence or absence of the disease as our label.

ROC and AUC in Decision Trees - Practical Example

- Suppose we're analyzing data from a medical study to predict the risk of a certain disease. We'll use one key health metric as our feature and the presence or absence of the disease as our label.
- Feature: "Blood Pressure Level" (a continuous scale from 0.1 to 1.0, where higher values indicate higher blood pressure).

ROC and AUC in Decision Trees - Practical Example

- Suppose we're analyzing data from a medical study to predict the risk of a certain disease. We'll use one key health metric as our feature and the presence or absence of the disease as our label.
- Feature: "Blood Pressure Level" (a continuous scale from 0.1 to 1.0, where higher values indicate higher blood pressure).
- Label: Diagnosis of a specific disease (0 = No disease, 1 = Disease present).

ROC and AUC in Decision Trees - Practical Example

Table: Patient Data

Blood Pressure Level	Disease Diagnosis
0.10	0
0.40	0
0.35	1
0.80	1
0.70	1
0.30	0
0.60	1
0.55	1
0.50	0
0.45	0

ROC and AUC in Decision Trees - Practical Example

- To demonstrate the calculation of a specific point on the ROC curve, let's choose a threshold and calculate the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) based on our dataset.

ROC and AUC in Decision Trees - Practical Example

- To demonstrate the calculation of a specific point on the ROC curve, let's choose a threshold and calculate the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) based on our dataset.
- Let's choose a threshold of 0.5 for the "Blood Pressure Level" to calculate a point on the ROC curve. This means we classify a patient as having the disease (positive) if their blood pressure level is equal to or higher than 0.5.

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.

Formulas:

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.
- False Positives (FP): Count of patients who do not have the disease (Label $= 0$) but their blood pressure level is ≥ 0.5.

Formulas:

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.
- False Positives (FP): Count of patients who do not have the disease (Label $= 0$) but their blood pressure level is ≥ 0.5.
- True Negatives (TN): Count of patients who do not have the disease (Label $= 0$) and their blood pressure level is < 0.5.

Formulas:

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.
- False Positives (FP): Count of patients who do not have the disease (Label $= 0$) but their blood pressure level is ≥ 0.5.
- True Negatives (TN): Count of patients who do not have the disease (Label $= 0$) and their blood pressure level is < 0.5.
- False Negatives (FN): Count of patients who have the disease (Label $= 1$) but their blood pressure level is < 0.5.

Formulas:

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.
- False Positives (FP): Count of patients who do not have the disease (Label $= 0$) but their blood pressure level is ≥ 0.5.
- True Negatives (TN): Count of patients who do not have the disease (Label $= 0$) and their blood pressure level is < 0.5.
- False Negatives (FN): Count of patients who have the disease (Label $= 1$) but their blood pressure level is < 0.5.

Formulas:

- $\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) =$??

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.
- False Positives (FP): Count of patients who do not have the disease (Label $= 0$) but their blood pressure level is ≥ 0.5.
- True Negatives (TN): Count of patients who do not have the disease (Label $= 0$) and their blood pressure level is < 0.5.
- False Negatives (FN): Count of patients who have the disease (Label $= 1$) but their blood pressure level is < 0.5.

Formulas:

- $\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) =$??
- $\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) =$??

ROC and AUC in Decision Trees - Practical Example

Calculation Process:

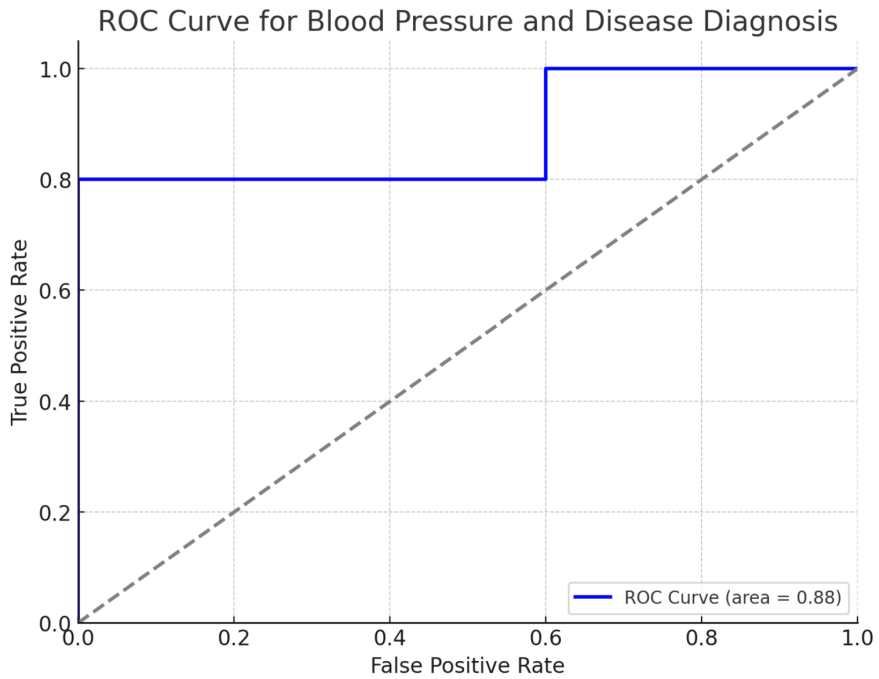
- True Positives (TP): Count of patients who have the disease (Label $=$ 1) and their blood pressure level is ≥ 0.5.
- False Positives (FP): Count of patients who do not have the disease (Label $= 0$) but their blood pressure level is ≥ 0.5.
- True Negatives (TN): Count of patients who do not have the disease (Label $= 0$) and their blood pressure level is < 0.5.
- False Negatives (FN): Count of patients who have the disease (Label $= 1$) but their blood pressure level is < 0.5.

Formulas:

- $\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) = 0.8$
- $\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) = 0.2$

ROC and AUC in Decision Trees - Practical Example

- By calculating all the points, we get the ROC Curve.



ROC and AUC in Decision Trees - Practical Example

- By calculating all the points, we get the ROC Curve.
- Verify AOC = 0.88 = 1-0.6*0.2