Business Intelligence

TICS-423
Universidad Adolfo Ibáñez

Week 05: 30 August - 02 September, 2016

Claudio Diaz
Sebastián Moreno
Gonzalo Ruz

Overview

- **Task Specifications:**
  - Exploratory Data Analysis
  - **Predictive Modeling**
  - Descriptive Modeling
  - Pattern Discovery
- **Data Representation:**
- **Knowledge representation**
- **Learning technique**
  - **Search + Scoring**
- **Prediction and/or interpretation**

Predictive modelling

Predictive modeling

- **Task Specification: Predictive Modeling**
- **Data Representation: Homogeneous IID data**
- **Knowledge representation**
- **Learning technique**
  - **Search + Scoring**
- **Prediction and/or interpretation**

Predictive modeling

- Predictive models predict the value of one variable of interest given known values of other variables
- Focus on modeling the conditional distribution P(Y | X) or on modeling the decision boundary for Y

Predictive modeling

- Predictive models predict the value of one variable of interest given known values of other variables
- Focus on modeling the conditional distribution P(Y | X) or on modeling the decision boundary for Y
- **Task:** estimate a predictive function $f(x;\theta)=y$
  - Assume that there is a function $y=f(x)$ that maps data instances (x) to class labels (y)
  - Construct a model that approximates the mapping
    Classification: if y is categorical
    Regression: if y is real-valued

Predictive modeling

- **Task:** estimate a predictive function $f(x;\theta)=y$
- **Data representation:** Training set: Paired attribute vectors and class labels $\langle y(i), x(i) \rangle$ or $n \times p$ tabular data with class label (y) and p-1 attributes (x)
- **Knowledge representation:** Underlying structure of the model or patterns that we seek from the data (naive bayes, decision tree, regression).
- **Learning:** Defines a space of possible models M

  Scoring function: Evaluate possible models to determine the model which best fits the data

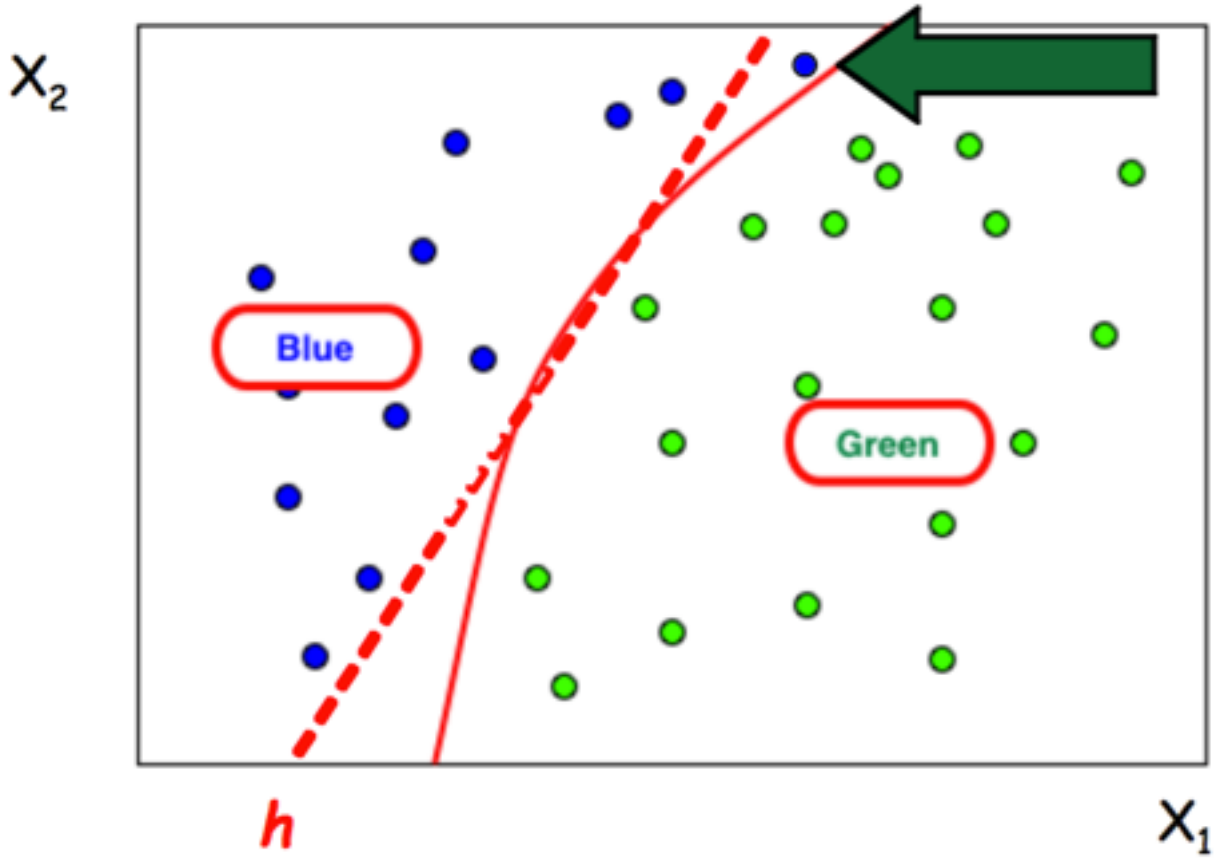
  Search: Search the space of models.

Predictive modelling
Modeling approach

PM: Modeling approaches

- **Classification:** In its simplest form, a classification model defines a decision boundary (h) and labels for each side of the boundary.

- **Input: x={x_1,x_2,…,x_n}** is a set of attributes, function f assigns a label y to input **x**, where y is a discrete variable with a finite number of values.



PM: Modeling approaches

- **Data:** decision-making process to go outside for recess to play.

Day	outlook	Temp	Humidity	Windy	Play
1	sunny	85	85	no	**NO**
2	sunny	80	90	yes	**NO**
3	overcast	83	86	no	**YES**
4	rainy	70	96	no	**YES**
5	rainy	68	80	no	**YES**
6	rainy	65	70	yes	**NO**
7	overcast	64	65	yes	**YES**
8	sunny	72	95	no	**NO**

PM: Modeling approaches

- **Classification output:** Different classification tasks can require different kinds of output.
    - **Class labels:** Each instance is assigned a single label. Model only need to decide on crisp class boundaries

Day	outlook	Temp	Humidity	Windy	Play	class label	ranking	probability
1	sunny	85	85	no	**NO**	**NO**		
2	sunny	80	90	yes	**NO**	**YES**		
3	overcast	83	86	no	**YES**	**YES**		
4	rainy	70	96	no	**YES**	**YES**		
5	rainy	68	80	no	**YES**	**NO**		
6	rainy	65	70	yes	**NO**	**YES**		
7	overcast	64	65	yes	**YES**	**YES**		
8	sunny	72	95	no	**NO**	**NO**		

PM: Modeling approaches

- **Classification output:** Different classification tasks can require different kinds of output.
  - **Ranking:** Instances are ranked according to their likelihood of belonging to a particular class
    Model implicitly explores many potential class boundaries

Day	outlook	Temp	Humidity	Windy	Play	class label	ranking	probability
5	rainy	68	80	no	**YES**	**NO**	**2,74**	
7	overcast	64	65	yes	**YES**	**YES**	**2,39**	
3	overcast	83	86	no	**YES**	**YES**	**2,07**	
6	rainy	65	70	yes	**NO**	**YES**	**1,82**	
4	rainy	70	96	no	**YES**	**YES**	**0,95**	
1	sunny	85	85	no	**NO**	**NO**	**-0,08**	
8	sunny	72	95	no	**NO**	**NO**	**-1,20**	
2	sunny	80	90	yes	**NO**	**YES**	**-2,68**	

PM: Modeling approaches

- **Classification output:** Different classification tasks can require different kinds of output.
  - **Probabilities**: Instances are assigned class probabilities p(y|**x**)
    Allows for more refined reasoning about sets of instances

Day	outlook	Temp	Humidity	Windy	Play	class label	ranking	probability
1	sunny	85	85	no	**NO**	**NO**	**-0,08**	**0,13**
2	sunny	80	90	yes	**NO**	**YES**	**-2,68**	**0,05**
3	overcast	83	86	no	**YES**	**YES**	**2,07**	**0,96**
4	rainy	70	96	no	**YES**	**YES**	**0,95**	**0,45**
5	rainy	68	80	no	**YES**	**NO**	**2,74**	**0,87**
6	rainy	65	70	yes	**NO**	**YES**	**1,82**	**0,77**
7	overcast	64	65	yes	**YES**	**YES**	**2,39**	**0,68**
8	sunny	72	95	no	**NO**	**NO**	**-1,20**	**0,16**

PM: Modeling approaches: Discriminative classification

- **Discriminative classification:**
  - Model the decision boundary directly
  - Direct mapping from inputs **x** to class label y
  - No attempt to model probability distributions
  - May seek a discriminant function f(**x**;θ) that maximizes measure of separation between classes.
  - Examples: Perceptrons, nearest neighbor classifiers, support vector machines, decision trees.

PM: Modeling approaches: Probabilistic classification

- **Probabilistic classification:**
  - Model the underlying probability distributions
  - Posterior class probabilities: $p(y|\mathbf{x})$
  - Class-conditional and class prior: $p(\mathbf{x}|y)$ and $p(y)$
  - Maps from inputs x to class label y indirectly through posterior class distribution $p(y|\mathbf{x})$
  - Examples: Naive Bayes classifier, logistic regression, probability estimation trees.

Predictive modelling
Knowledge representation

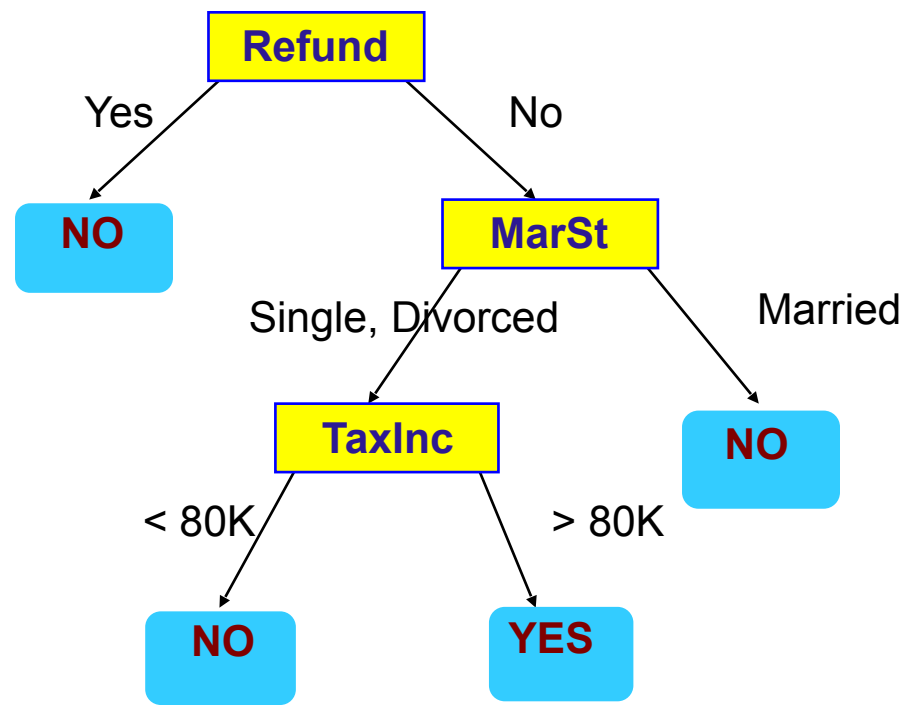
PM: Knowledge representation

- Underlying structure of the model or patterns that we seek from the data
    - Defines space of possible models for algorithm to search over
- **Model**: high-level global description of dataset
    - “All models are wrong, some models are useful”
      G. Box and N. Draper (1987)
    - **Choice of model family determines space of parameters and structure**
    - **Estimate model parameters and possibly model structure from training data**

PM: Knowledge representation

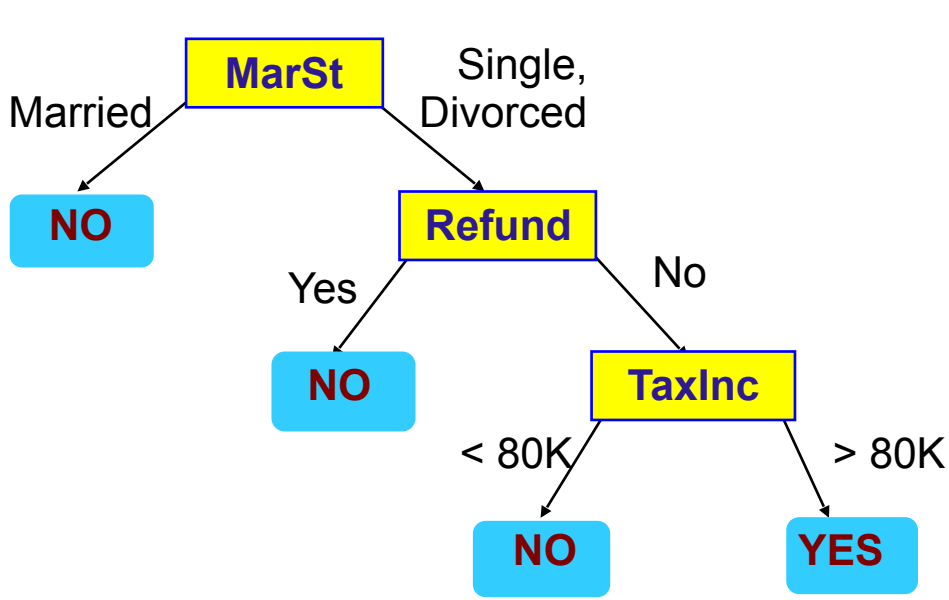
- **Model:** decision tree

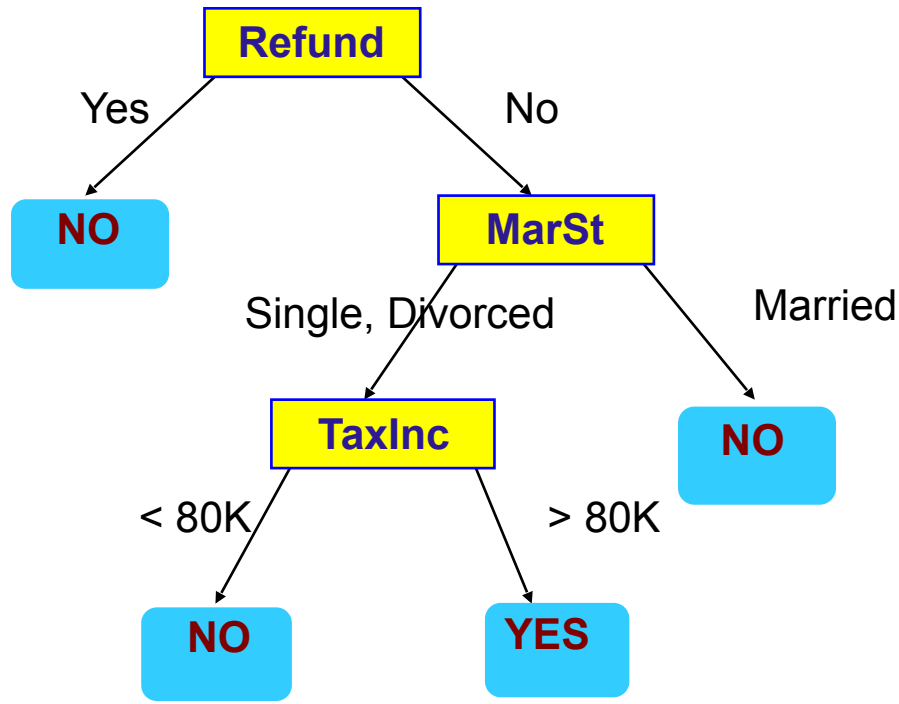
	categorical	categorical	continuous	class
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



PM: Knowledge representation

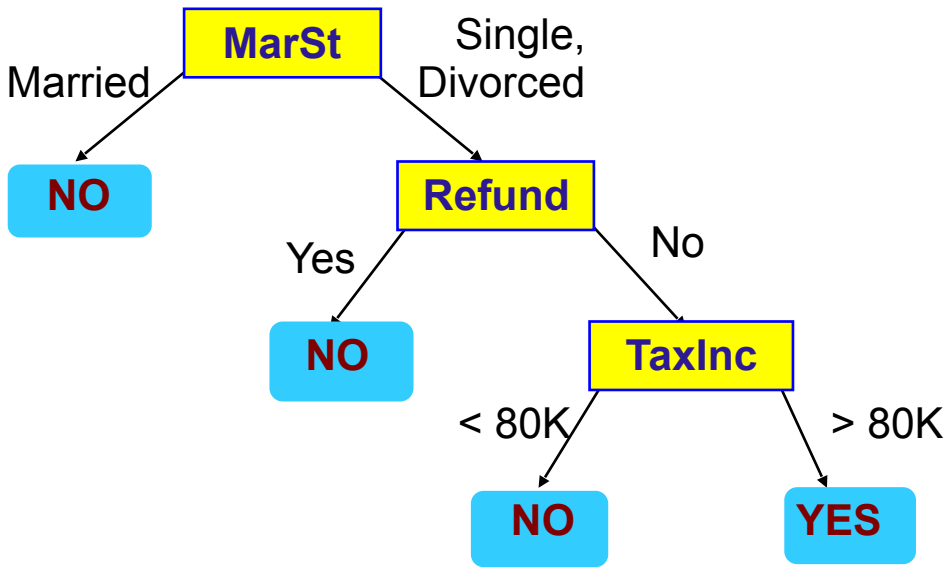
- **Model:** decision tree
- **Model space:** All possible decision trees

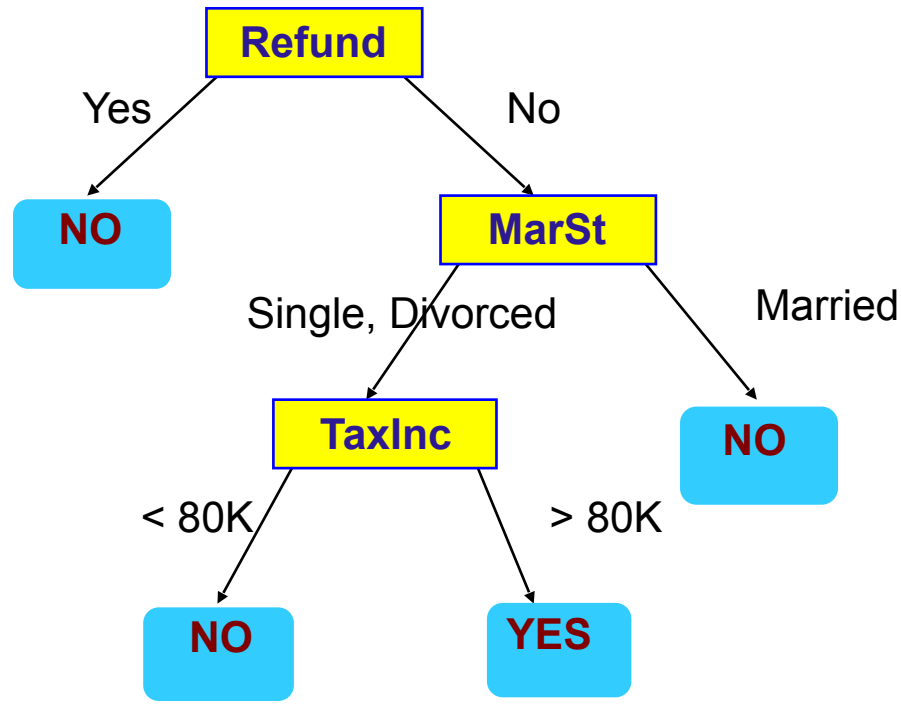




PM: Knowledge representation

- **Model:** decision tree
- **Model space:** All possible decision trees
- To **learn** the model (search the best model over the model space), we need to define the **scoring function** and the **search procedure.**





PM: Knowledge representation

- **Model:** decision rule
- **Model space:** all possible rules formed from conjunctions of features

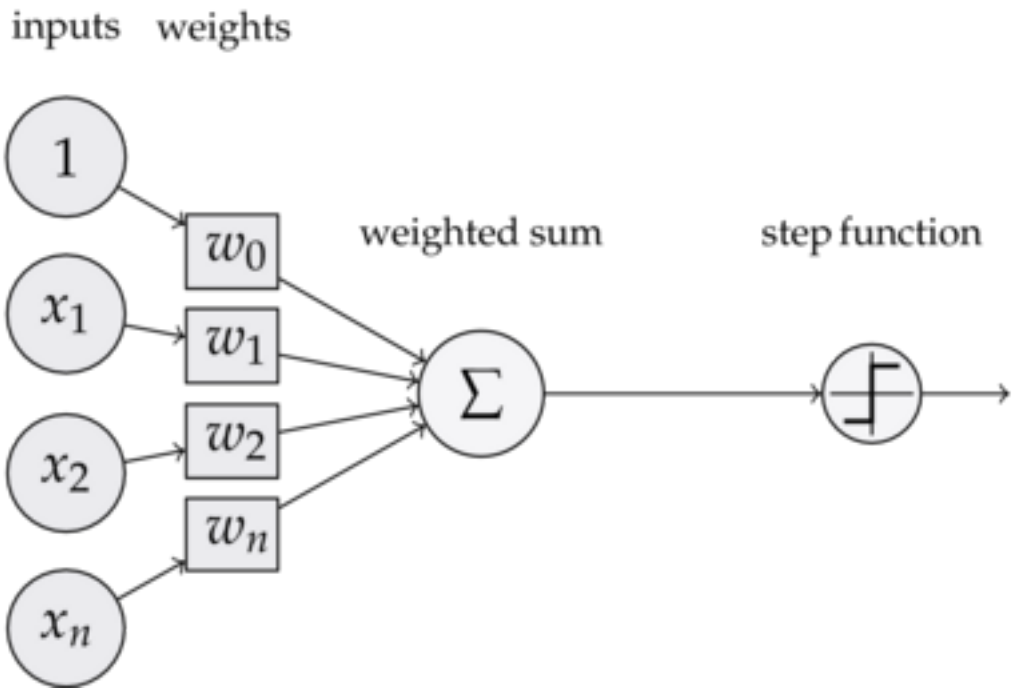
TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

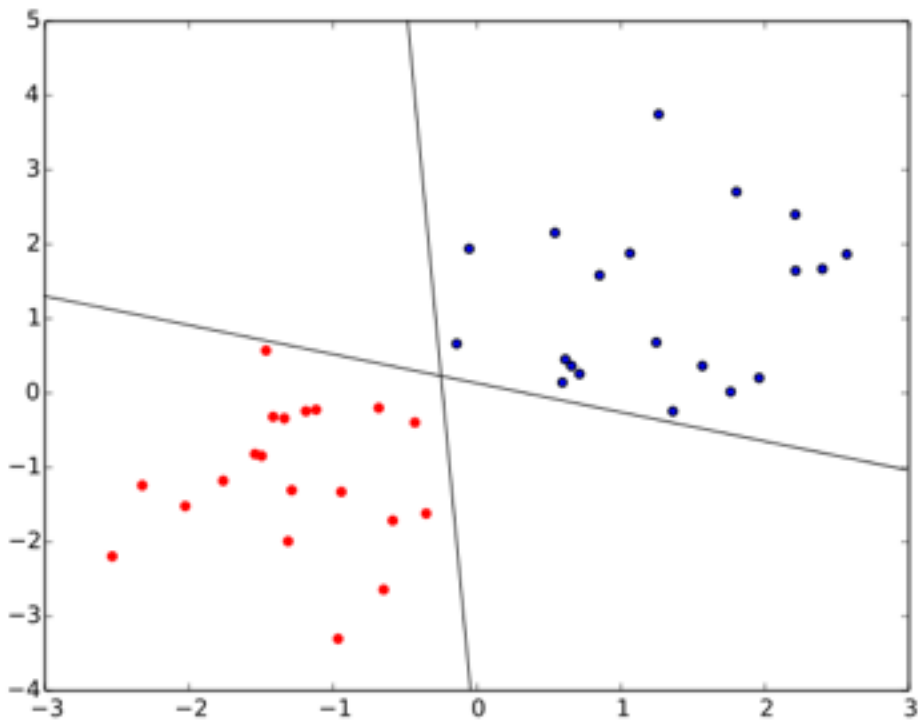
Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

PM: Knowledge representation

- **Model:** perceptron
- **Model space:** weights w, for each of j attributes

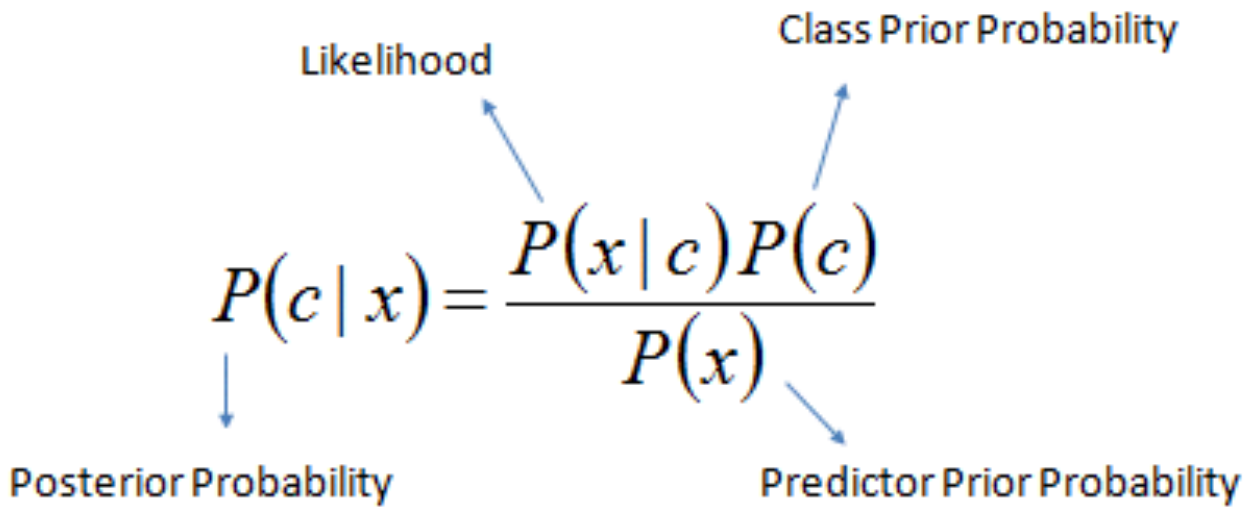




$$f(x) = \begin{cases} 1 & \sum w_j x_j > 0 \\ 0 & \sum w_j x_j \leq 0 \end{cases}$$

PM: Knowledge representation

- **Model:** naive bayes
- **Model space:**
  parameters in conditional distributions p(xi|y)
  parameters in prior distribution p(y)

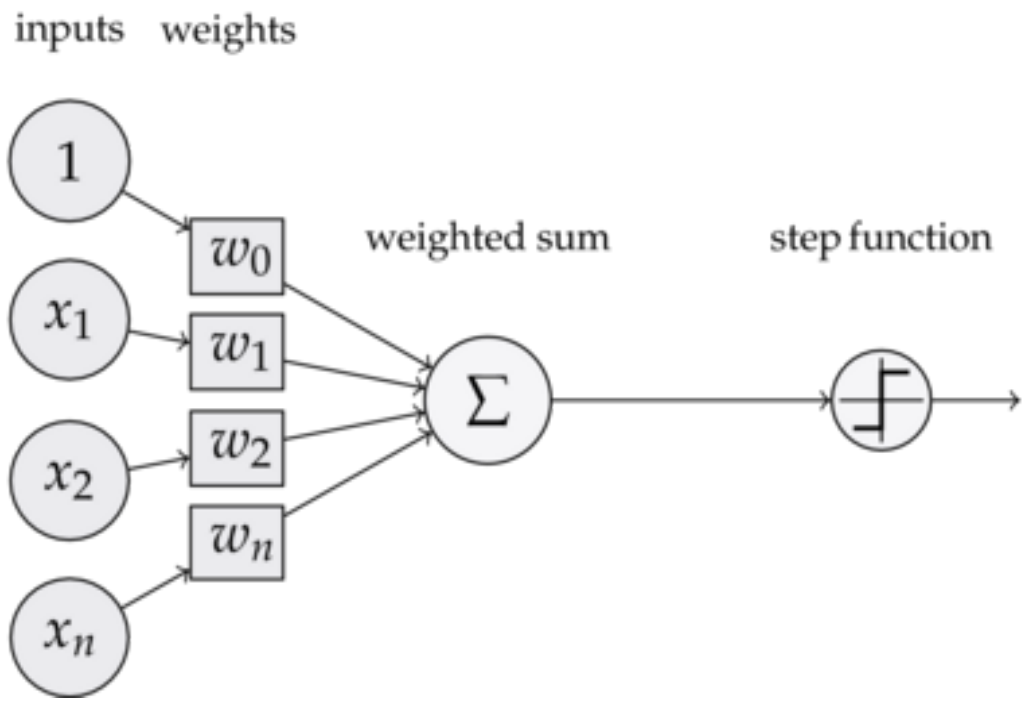


$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

PM: Knowledge representation

- **Parametric vs. non-parametric models**

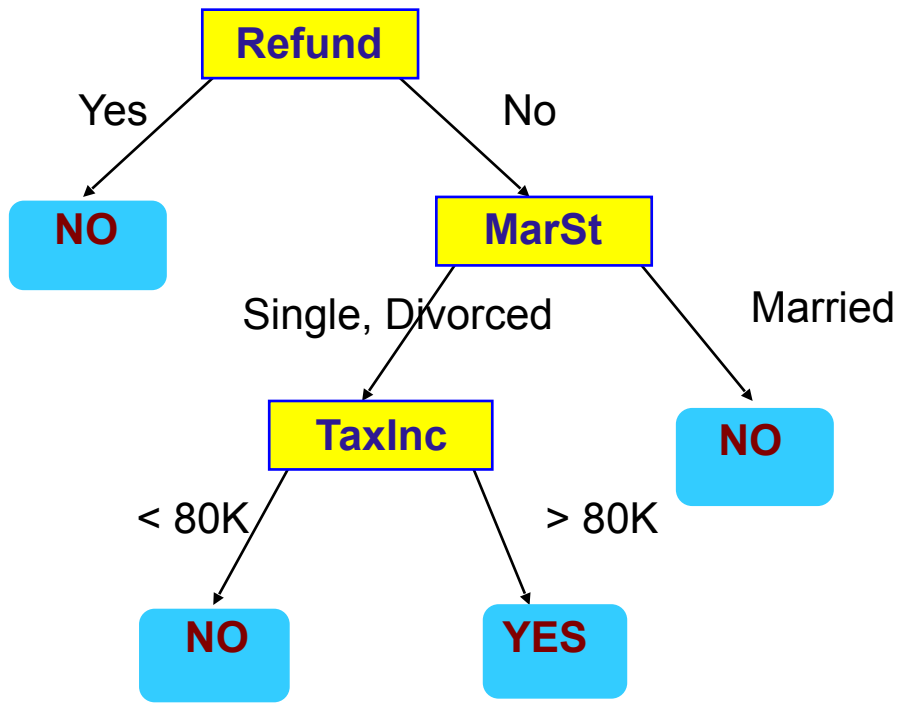
- **Parametric:**
  Particular functional form is assumed
  Number of parameters is fixed in advance
  Examples: Naive Bayes, perceptron



PM: Knowledge representation

- **Parametric vs. non-parametric models**

- **Non-parametric:**
  Few assumptions are made about the functional form
  Model structure is determined from data
  Examples: classification tree, nearest neighbor



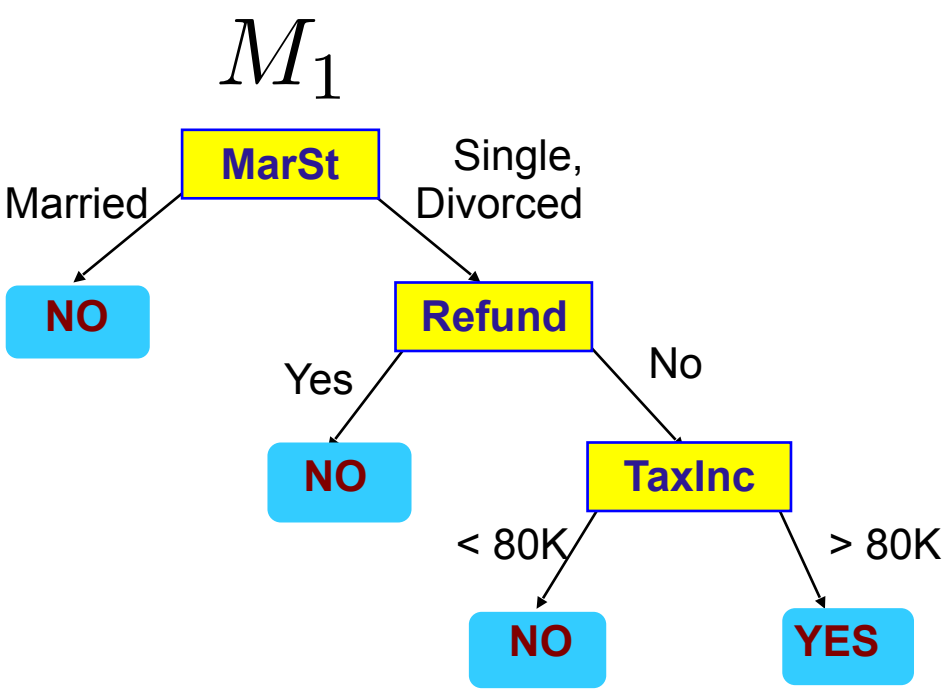
Predictive modelling
Learning

PM: Learning

- **Learning predictive models:**
  - Choose a **data representation**
  - Select a **knowledge representation** (a “model”)
    - Defines a space of possible models **M**={M_1, M_2, ..., M_k}

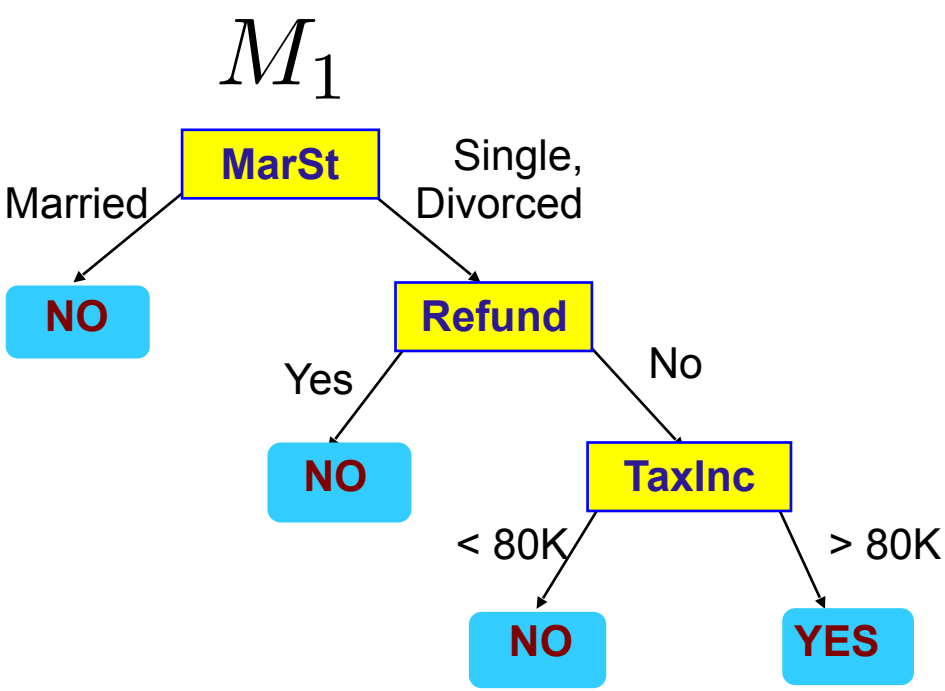
PM: Learning

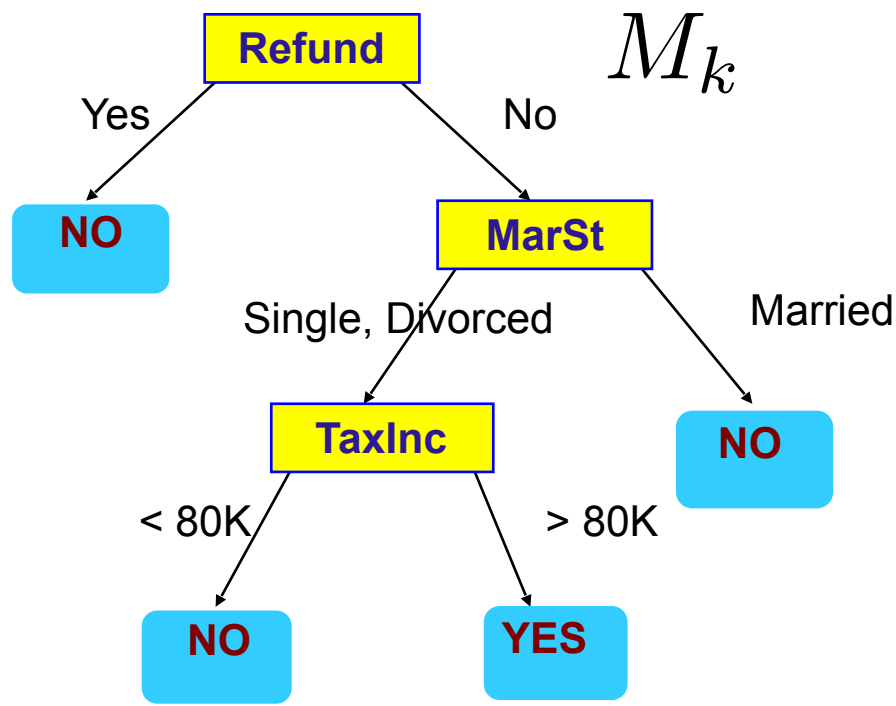
- **Learning predictive models:**
  - Choose a **data representation**
  - Select a **knowledge representation** (a “model”)
    - Defines a space of possible models **M**={M_1, M_2, ..., M_k}



PM: Learning

- **Learning predictive models:**
  - Choose a **data representation**
  - Select a **knowledge representation** (a “model”)
    - Defines a space of possible models **M**={M_1, M_2, ..., M_k}





PM: Learning

- **Learning predictive models:**
  - Choose a **data representation**
  - Select a **knowledge representation** (a “model”)
    - Defines a space of possible models **M**={M_1, M_2, ..., M_k}
  - Use **search** to identify “best” model(s)
    - Search the space of models (i.e., with alternative structures and/or parameters)
    - Evaluate possible models with **scoring function** to determine the model which best fits the data

PM: Learning

- **Learning predictive models:**
  - Choose a **data representation**
  - Select a **knowledge representation** (a “model”)
    - Defines a space of possible models **M**={M_1, M_2, ..., M_k}
  - Use **search** to identify “best” model(s)
    - Search the space of models (i.e., with alternative structures and/or parameters)
    - Evaluate possible models with **scoring function** to determine the model which best fits the data

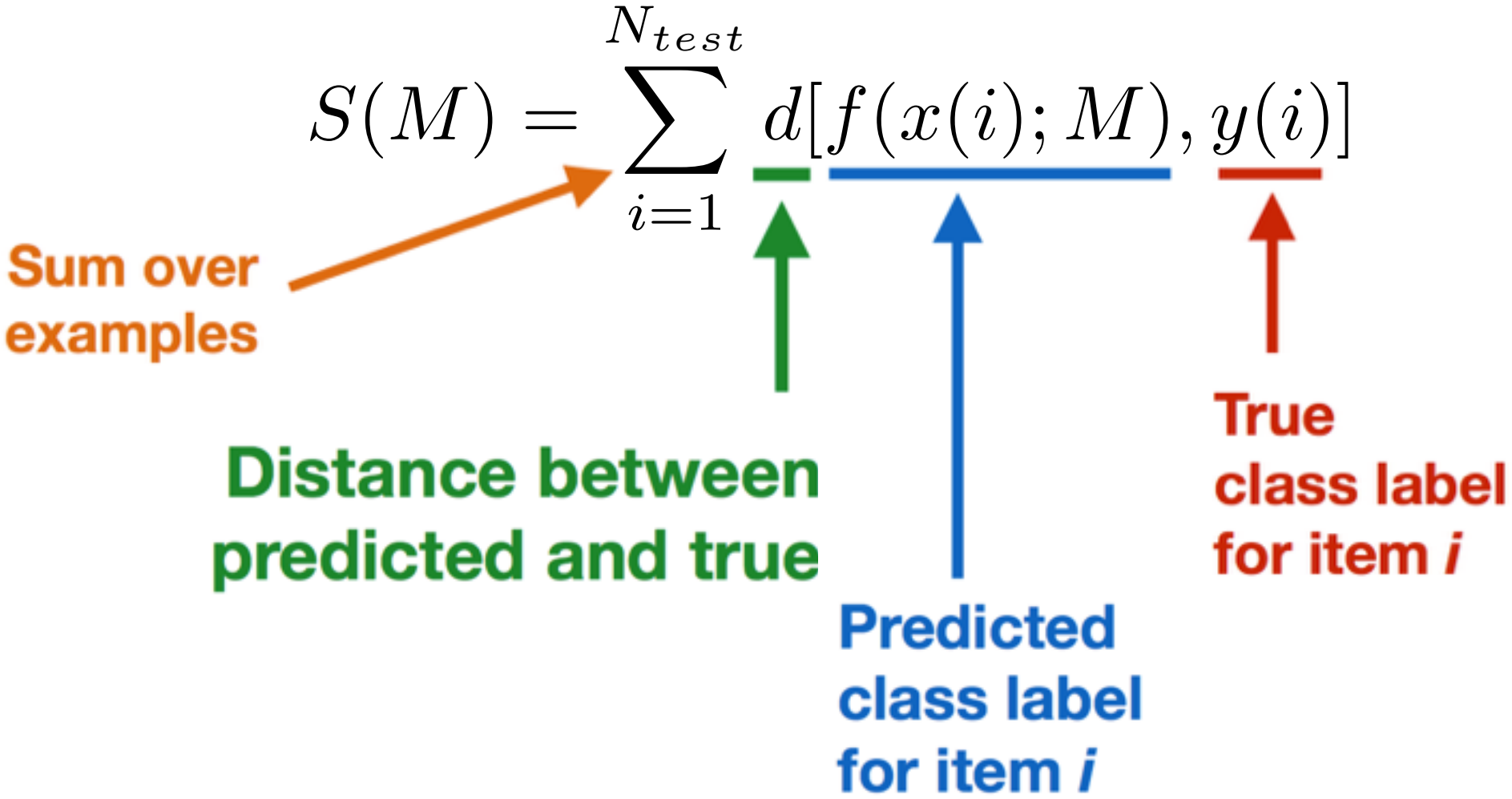
$$score(M_1) = 3.67 \qquad score(M_k) = 2.49$$

PM: Learning, scoring function

- Given a model **M** and dataset **D**, we would like to "score" model **M** with respect to **D**
  - Assess the quality of predictions for a set of instances: Measures difference between the prediction **M** makes for an instance i and the true class label value of i
  - Goal is to rank the models in terms of their utility (for capturing **D**) and choose the "best" model
  - Score function can be used to search over **parameters** and/or **model structure**

PM: Learning, scoring function

- General scoring function:



PM: Learning, scoring function

- Common score functions:
  - Zero-one loss: $$S_{0/1}(M) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} I[f(x(i); M), y(i)]$$

$$\text{where } I(a,b) = \begin{cases} 1 & a \neq b \\ 0 & \text{otherwise} \end{cases}$$

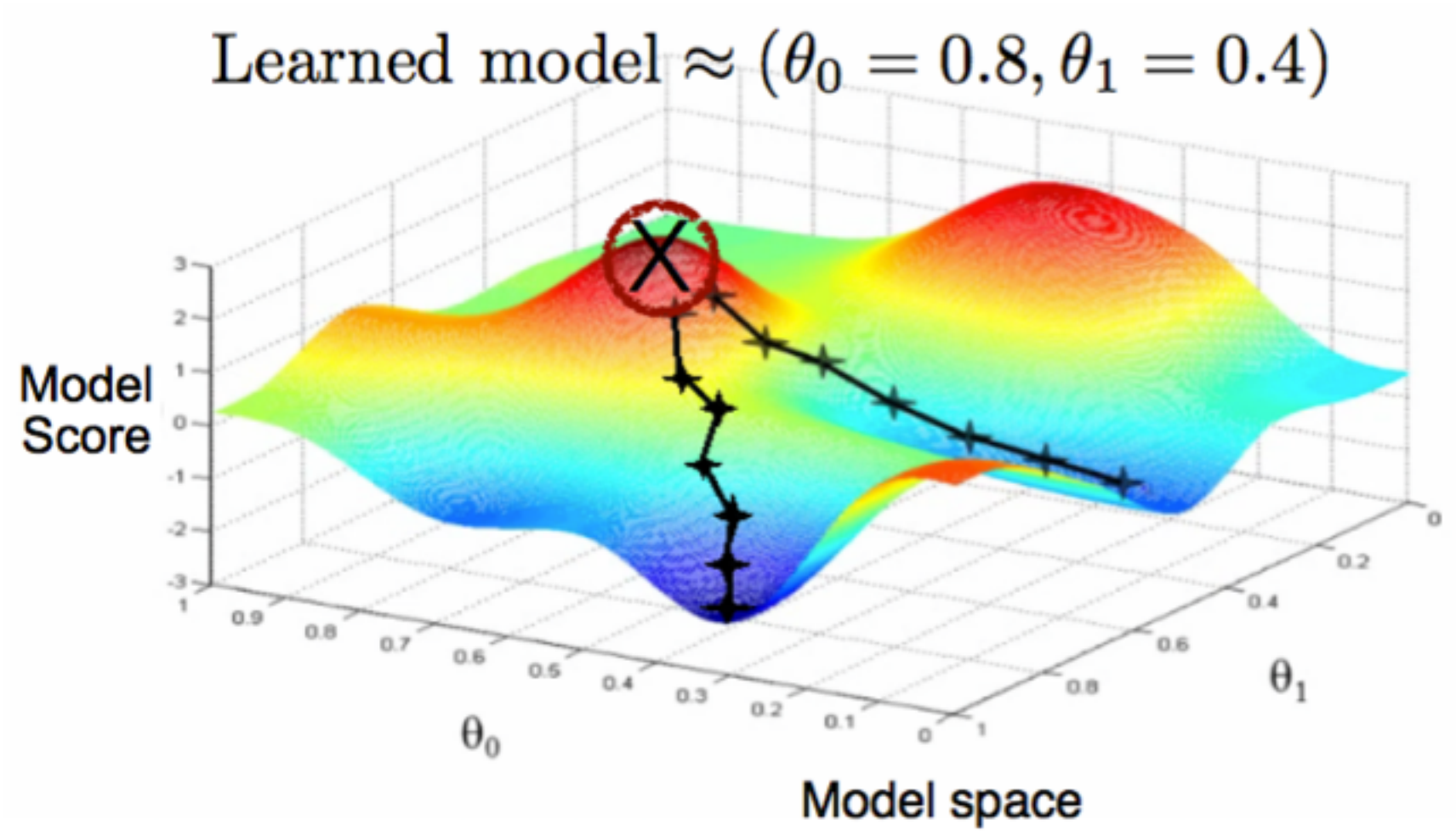
  - Squared loss: $$S_{sq}(M) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} [f(x(i); M) - y(i)]^2$$

PM: Learning, search

- Consider a space of possible models $\mathbf{M}=\{M_1, M_2, \ldots, M_K\}$ with parameters $\boldsymbol{\theta}$
- Search could be over model structures or parameters, example:
  - Parameters: In a linear regression model, find the regression coefficients ($\boldsymbol{\beta}$) that minimize squared loss on the training data.
  - Model structure: In a decision trees, find the tree structure that maximizes accuracy on the training data.

PM: Learning, search

- Example: Search for the set of parameters **θ** that maximize the model score.



Predictive modelling
Naive Bayes

Predictive modeling, naive bayes

- Naive bayes learns a conditional probability distribution.
- Given a data point **x**, the output of the model is the probability of **x** belonging to a specific class.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- How likely is to buy a computer a 23 years old student with a fair credit and medium income?
  x={<=30,medium,yes,fair}
- **P(BC=yes|A<=30; I=med; S=yes; CR=fair) ∝ 0.028**
- **P(BC=no|A<=30; I=med; S=yes; CR=fair) ∝ 0.007**

Predictive modeling, naive bayes

- Naive bayes uses 3 key aspects: conditional probability, bayesian theorem, and conditional independence.
- Conditional probability $P(\mathbf{X}|C) = \frac{P(\mathbf{X}, C)}{P(C)} \qquad P(C|\mathbf{X}) = \frac{P(\mathbf{X}, C)}{P(\mathbf{X})}$
- Bayesian theorem: $P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})}$
- Conditional independence: If X_1, X_2, …, X_k are independent given C then

$$P(X_1, X_2, \ldots, X_k|C) = \prod_{i=1}^{k} P(X_i|C)$$

Predictive modeling, naive bayes

- Bayesian theorem example

- **Given that:**
  meningitis produces torticollis 50% of the times => P(T|M)=0.5
  The probability of meningitis is 1/50,000 => P(M)=1/50000
  The probability of torticollis is 1/20 => P(T)=1/20

- If a patient has torticollis, what is probability of meningitis?

$$P(M|T) = \frac{P(T|M)P(M)}{P(T)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Predictive modeling, naive bayes

- Assume each attribute and class as variables
- Given a dataset with attributes (X_1, X_2,…,X_k)
- The objective is to predict the class C
- Specifically, we want to find the model that maximize P(C|X_1, X_2,…,X_k) given all the points of the dataset (training)
- However, can we estimate P(C|X_1, X_2,…,X_k) from the data?

$$P(C|X_1, X_2, \ldots, X_k) = \frac{P(X_1, X_2, \ldots, X_k|C)P(C)}{P(X_1, X_2, \ldots, X_k)}$$

Predictive modeling, naive bayes

$$P(C|X_1, X_2, \ldots, X_k) = \frac{P(X_1, X_2, \ldots, X_k|C)P(C)}{P(X_1, X_2, \ldots, X_k)}$$

- Given the data **D**, P(C) is easy to calculate. This distribution is the number of elements of each class over the total number of data points.
- How can we compute $P(X_1, X_2,\ldots,X_k)$? It is **complicate** and **unnecessary**, because $P(X_1, X_2,\ldots,X_k)$ is a normalizing factor to make probabilities sum to 1.
- Example, to compare $P(C=c_1|X_1=x_1,\ldots,X_k=x_k)$ and $P(C=c_2|X_1=x_1,\ldots,X_k=x_k)$

$$\frac{P(x_1, x_2, \ldots, x_k|c_1)P(c_1)}{P(x_1, x_2, \ldots, x_k)} \quad \frac{P(x_1, x_2, \ldots, x_k|c_2)P(c_2)}{P(x_1, x_2, \ldots, x_k)}$$

Predictive modeling, naive bayes

$$P(C|X_1, X_2, \ldots, X_k) \propto P(X_1, X_2, \ldots, X_k|C)P(C)$$

- How can we compute $P(X_1, X_2, \ldots, X_k|C)$? It is **complicate** and **necessary**.
- **Trick**: **NAIVELY ASSUME** conditional independence of $X_1, X_2, \ldots, X_k$ given C (even though this is not necessarily true).

$$P(C|X_1, X_2, \ldots, X_k) \propto \prod_{i=1}^{k} P(X_i|C)P(C)$$

- Now we need to compute each $P(X_i|C)$ from the data.
- A new point is classified as C_j if $P(C_j|\mathbf{X})$ is the maximum among all $P(C_i|\mathbf{X})$

Predictive modeling, naive bayes, learning

$$P(BC|A,I,S,CR) = \frac{P(A,I,S,CR|BC)P(BC)}{P(A,I,S,CR)}$$
$$= \frac{P(A|BC)P(I|BC)P(S|BC)P(CR|BC)P(BC)}{P(A,I,S,CR)}$$
$$\propto P(A|BC)P(I|BC)P(S|BC)P(CR|BC)P(BC)$$

NBC parameters = CPDs+prior

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

CPDs: P(A|BC)
P(I|BC)
P(S|BC)
P(CR|BC)
Prior: P(BC)

Predictive modeling, naive bayes, learning

- The **scoring** function is the likelihood function L(**θ**|**X**)=P(**X**|**θ**), where **θ** are the probability distributions. The **searching procedure** is the maximum likelihood estimation, which analytically estimates the best parameters for **θ**

- Let **D** a dataset with n points, k attributes, and class labels 1 to m
  Let $|C_i|$ be the number of elements belonging to class i
  Let $|X_{jk}|$ be the number of elements from attribute j with value k

- Prior: $P(C = i) = \frac{|C_i|}{n}$

- Discrete attributes: $P(X_j = k|C = i) = \frac{|X_{jk}|}{|C_i|}$

- Continuous attributes:
  option 1) Discretization of the attributes in the desired number of ranges.
  option 2) Assume a continuous distribution for the data (usually normal).

Predictive modeling, naive bayes, example

age	income	student	credit	buys
<=30	high	no	fair	**no**
<=30	high	no	excellent	**no**
31…40	high	no	fair	**yes**
>40	medium	no	fair	**yes**
>40	low	yes	fair	**yes**
>40	low	yes	excellent	**no**
31…40	low	yes	excellent	**yes**
<=30	medium	no	fair	**no**
<=30	low	yes	fair	**yes**
>40	medium	yes	fair	**yes**
<=30	medium	yes	excellent	**yes**
31…40	medium	no	excellent	**yes**
31…40	high	yes	fair	**yes**
>40	medium	no	excellent	**no**

- Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

P(BC)

BC	
YES	
NO	

Predictive modeling, naive bayes, example

age	income	student	credit	buys	
<=30	high	no	fair	**no**	**1**
<=30	high	no	excellent	**no**	**2**
31…40	high	no	fair	**yes**	**3**
>40	medium	no	fair	**yes**	**4**
>40	low	yes	fair	**yes**	**5**
>40	low	yes	excellent	**no**	**6**
31…40	low	yes	excellent	**yes**	**7**
<=30	medium	no	fair	**no**	**8**
<=30	low	yes	fair	**yes**	**9**
>40	medium	yes	fair	**yes**	**10**
<=30	medium	yes	excellent	**yes**	**11**
31…40	medium	no	excellent	**yes**	**12**
31…40	high	yes	fair	**yes**	**13**
>40	medium	no	excellent	**no**	**14**

- Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

n=14

P(BC)

BC	
YES	
NO	

Predictive modeling, naive bayes, example

age	income	student	credit	buys	
<=30	high	no	fair	**no**	
<=30	high	no	excellent	**no**	
31…40	high	no	fair	**yes**	**1**
>40	medium	no	fair	**yes**	**2**
>40	low	yes	fair	**yes**	**3**
>40	low	yes	excellent	**no**	
31…40	low	yes	excellent	**yes**	**4**
<=30	medium	no	fair	**no**	
<=30	low	yes	fair	**yes**	**5**
>40	medium	yes	fair	**yes**	**6**
<=30	medium	yes	excellent	**yes**	**7**
31…40	medium	no	excellent	**yes**	**8**
31…40	high	yes	fair	**yes**	**9**
>40	medium	no	excellent	**no**	

- Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

BC=yes => 9

P(BC)

BC	
YES	**9/14**
NO	

Predictive modeling, naive bayes, example

age	income	student	credit	buys	
<=30	high	no	fair	**no**	**1**
<=30	high	no	excellent	**no**	**2**
31…40	high	no	fair	**yes**	
>40	medium	no	fair	**yes**	
>40	low	yes	fair	**yes**	
>40	low	yes	excellent	**no**	**3**
31…40	low	yes	excellent	**yes**	
<=30	medium	no	fair	**no**	**4**
<=30	low	yes	fair	**yes**	
>40	medium	yes	fair	**yes**	
<=30	medium	yes	excellent	**yes**	
31…40	medium	no	excellent	**yes**	
31…40	high	yes	fair	**yes**	
>40	medium	no	excellent	**no**	**5**

- Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

BC=no => 5

P(BC)

BC	
YES	**9/14**
NO	**5/14**

Predictive modeling, naive bayes, example

age	income	student	credit	buys
<=30	high	no	fair	**no**
<=30	high	no	excellent	**no**
31…40	high	no	fair	**yes**
>40	medium	no	fair	**yes**
>40	low	yes	fair	**yes**
>40	low	yes	excellent	**no**
31…40	low	yes	excellent	**yes**
<=30	medium	no	fair	**no**
<=30	low	yes	fair	**yes**
>40	medium	yes	fair	**yes**
<=30	medium	yes	excellent	**yes**
31…40	medium	no	excellent	**yes**
31…40	high	yes	fair	**yes**
>40	medium	no	excellent	**no**

- Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

P(A|BC)

BC	A	
YES	<=30	**2/9**
	31..40	**4/9**
	>40	**3/9**
NO	<=30	
	31..40	
	>40	

Predictive modeling, naive bayes, example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- Estimate prior P(BC) and conditional probability distributions P(A | BC), P(I | BC), P(S | BC), P(CR | BC) independently with maximum likelihood estimation

P(A | BC)

BC	A	
YES	<=30	2/9
	31..40	4/9
	>40	3/9
NO	<=30	3/5
	31..40	0/5
	>40	2/5

P(I | BC)

BC	I	
YES	high	2/9
	med	4/9
	low	3/9
NO	high	2/5
	med	2/5
	low	1/5

P(S | BC)

BC	S	
YES	yes	6/9
	no	3/9
NO	yes	1/5
	no	4/5

P(CR | BC)

BC	CR	
YES	exc	3/9
	fair	6/9
NO	exc	4/5
	fair	2/5

P(BC)

BC	
YES	9/14
NO	5/14

Predictive modeling, naive bayes, example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- **PREDICTION**: How likely is to buy a computer a 23 years old student with a fair credit and medium income?

P(BC=yes|A<=30; I=med; S=yes; CR=fair)

$\propto$ P(A<=30|BC=y)P(I=med|BC=y)

P(S=y|BC=y)P(CR=fair|BC=y)P(BC=y)

= 2/9 * 4/9 * 6/9 * 6/9 * 9/14 ≈ 0.028

P(A | BC)

BC	A	
YES	**<=30**	**2/9**
	31..40	4/9
	>40	3/9
NO	<=30	3/5
	31..40	0/5
	>40	2/5

P(I | BC)

BC	I	
YES	high	2/9
	med	**4/9**
	low	3/9
NO	high	2/5
	med	2/5
	low	1/5

P(S | BC)

BC	S	
YES	**yes**	**6/9**
	no	3/9
NO	yes	1/5
	no	4/5

P(CR | BC)

BC	CR	
YES	exc	3/9
	fair	**6/9**
NO	exc	4/5
	fair	2/5

P(BC)

BC	
YES	**9/14**
NO	5/14

Predictive modeling, naive bayes, example

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- **PREDICTION**: How likely is to NOT buy a computer a 23 years old student with a fair credit and medium income?

P(BC=no|A<=30; I=med; S=yes; CR=fair)

$\propto$ P(A<=30|BC=n)P(I=med|BC=n)

P(S=y|BC=n)P(CR=fair|BC=n)P(BC=n)

= 3/5 * 2/5 * 1/5 * 2/5 * 5/14 ≈ 0.007

P(A | BC)

BC	A	
YES	<=30	2/9
	31..40	4/9
	>40	3/9
NO	**<=30**	**3/5**
	31..40	0/5
	>40	2/5

P(I | BC)

BC	I	
YES	high	2/9
	med	4/9
	low	3/9
NO	high	2/5
	med	**2/5**
	low	1/5

P(S | BC)

BC	S	
YES	yes	6/9
	no	3/9
NO	**yes**	**1/5**
	no	4/5

P(CR | BC)

BC	CR	
YES	exc	3/9
	fair	6/9
NO	exc	4/5
	fair	**2/5**

P(BC)

BC	
YES	9/14
NO	**5/14**

Predictive modeling, naive bayes, example

refund	status	income	affair
yes	single	125K	**no**
no	married	100K	**no**
no	single	70K	**no**
yes	married	120K	**no**
no	Divorced	95K	**yes**
no	married	60K	**no**
yes	Divorced	220K	**no**
no	single	85K	**yes**
no	married	75K	**no**
no	single	90K	**yes**

- To calculate P(income|affair) we have two options
  - Discretize the income and treated like a discrete parameter
  - Estimate a number of normal distributions equal to the number of classes
- Example: P(income|affair=**no)**, we obtain a normal distribution with mean 110K, and standard deviation 55.

$$P(\text{income} = 120K|\text{affair} = no) = \frac{1}{\sqrt{2\pi}55}\exp\left(-\frac{(120-110)^2}{2*55^2}\right) \approx 0.0072$$

Predictive modeling, naive bayes, smoothing

- **Zero counts are a problem**
- If an attribute value does not occur in training example, we assign zero probability to that value
- How does that affect the conditional probability P[f(**x**) | **x**] ? It equals 0!!!

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- How likely is to NOT buy a computer a 31 to 40 years old person with a fair credit and medium income?

P(BC=no|A<=31..40; I=med; S=no; CR=fair) = 0.

P(A|BC)

BC	A	
YES	<=30	2/9
	31..40	4/9
	>40	3/9
NO	<=30	3/5
	31..40	**0/5**
	>40	2/5

Predictive modeling, naive bayes, smoothing

- **Zero counts are a problem**
- To avoid the zero probabilities, we smooth the probability estimation by adding values in the numerator and denominator of the estimation.
- Laplace correction: Add 1 numerator and add |C| (the number of classes) to the denominator.

$$P(X_j = k|C = i) = \frac{|X_{jk}| + 1}{|C_i| + |C|}$$

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

P(A|BC)

BC	A	
YES	<=30	2/9
	31..40	4/9
	>40	3/9
NO	<=30	3/5
	31..40	**0/5**
	>40	2/5

P(A|BC)

BC	A	
YES	<=30	3/12
	31..40	5/12
	>40	4/12
NO	<=30	4/8
	31..40	**1/8**
	>40	3/8

Predictive modeling, naive bayes, summary

- **Strengths**:
  - Easy to implement and can be learned incrementally
  - Often performs well even when assumption is violated
  - Can be learned incrementally
  - Missing values are ignored in the learning process
  - Robust model with respect to outliers and irrelevant data
- **Weaknesses**:
  - Class conditional assumption produces skewed probability estimates
  - Dependencies among variables cannot be modeled

Predictive modeling, naive bayes, summary

- Task Specification: **Predictive Modeling**
- Data Representation: **Homogeneous IID data**
- Knowledge representation: **Naive Bayes**
  Model space: Parametric model with specific form which vary based on parameter estimates in CPDs
- Learning
  - Search algorithm: Maximum Likelihood Estimation optimization of parameters (convex optimization results in exact solution)
  - Scoring function: Likelihood of data given NBC model form
- Prediction: A proportional probability of the class given the data.