Bayesian decision theory

Nuno Vasconcelos

ECE Department, UCSD

Notation

- the notation in DHS is quite sloppy
  - e.g. show that

$$P(error) = \int P(error \mid z) P(z) dz$$

  - really not clear what this means
- we will use the following notation

$$\boxed{P_{X|Y}(x_0 \mid y_0)}$$

  - subscripts are random variables (uppercase)
  - arguments are the values of the random variables (lowercase)
  - equivalent to $$P(X = x_0 \mid Y = y_0)$$

Bayesian decision theory

- framework for computing optimal decisions on problems involving uncertainty (probabilities)
- basic concepts:
  - world:
    - has states or classes, drawn from a state or class random variable Y
    - fish classification, $Y \in \{bass, salmon\}$
    - student grading, $Y \in \{A, B, C, D, F\}$
    - medical diagnosis $\in$ {disease A, disease B, …, disease M}
  - observer:
    - measures observations (features), drawn from a random process X
    - fish classification, X = (scale length, scale width) $\in R^2$
    - student grading, $X = (HW_1, \ldots, HW_n) \in R^n$
    - medical diagnosis X = (symptom 1, …, symptom n) $\in R^n$

Bayesian decision theory

- decision function:
  - observer uses the observations to make decisions about the state of the world y
  - if x $\in \Omega$ and y $\in \Psi$ the decision function is the mapping

    $$g : \Omega \rightarrow \Psi$$

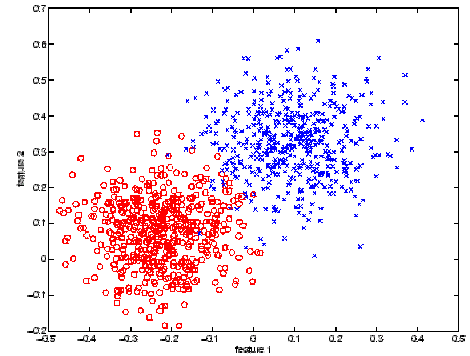
    such that

    $$g(x) = y_o$$

    and y_o is a prediction of the state y
- loss function:
  - is the cost *$L(y_o,y)$* of deciding for y_o when the true state is y
  - usually this is zero if there is no error and positive otherwise

- goal: to determine the optimal decision function for the loss *L(.,.)*

Classification



- we will focus on classification problems
  - the observer tries to infer the state of the world

$$g(x) = i, \quad i \in \{1, \ldots, M\}$$

  - we will also mostly consider the “0-1” loss function

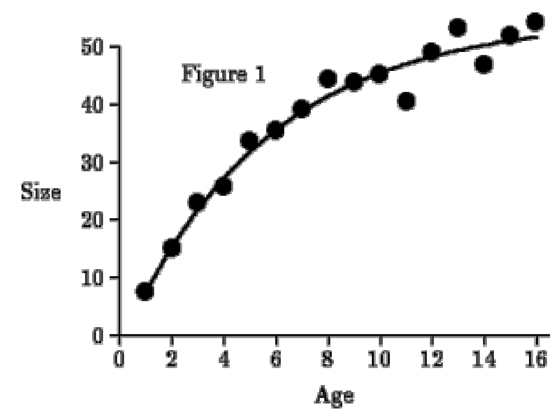
$$\boxed{L[g(x), y] = \begin{cases} 1, & g(x) \neq y \\ 0, & g(x) = y \end{cases}}$$

- but the regression case
  - the observer tries to predict a continuous y

$$g(x) \in \Re$$

  - is basically the same, for a suitable loss function, e.g. squared error

$$\boxed{L[g(x), y] = \|y - g(x)\|^2}$$



Tools for solving BDT problem

- probabilistic representations
  - joint distribution
  - class-conditional distributions
  - class probabilities

- properties of probabilistic inference
  - chain rule of probability
  - marginalization
  - independence
  - Bayes rule

Tools for solving BDT problem

- in order to find optimal decision function we need a probabilistic description of the problem
  - in the most general form this is the joint distribution

$$P_{X,Y}(x,i)$$

  - but we frequently decompose it into a combination of two terms

$$P_{X,Y}(x,i) = \underbrace{P_{X|Y}(x|i)}\underbrace{P_Y(i)}$$

  - these are the "class conditional distribution" and "class probability"
  - class probability
    - prior probability of state i, before observer actually measures anything
    - reflects a "prior belief" that, if all else is equal, the world will be in state i with probability $P_Y(i)$

Tools for solving BDT problem

- class-conditional distribution:
  - is the model for the observations given the class or state of the world
- consider the grading example
  - I know, from experience, that a% of the students will get A's, b% B's, c% C's, and so forth
  - hence, for any student, P(A) = a/100, P(B) = b / 100, etc.
  - these are the state probabilities, before I get to see any of the student's work
  - the class-conditional densities are the models for the grades themselves
  - let's assume that the grades are Gaussian, i.e. they are completely characterized by a mean and a variance

Tools for solving BDT problem

- knowledge of the class changes the mean grade, e.g. I expect
  - A students to have an average HW grade of 90%
  - B students 75%
  - C students 60%, etc
- this means that

$$P_{X|Y}(x|i) = G(x, \mu_i, \sigma)$$

- i.e. the distribution of class i is a Gaussian of mean μ_i and variance σ

▶ note that the decomposition

$$\boxed{P_{X,Y}(x,i) = P_{X|Y}(x|i)P_Y(i)}$$

is a special case of a very powerful tool in Bayesian inference

Tools for solving BDT problem

- probabilistic representations
  - joint distribution
  - class-conditional distributions
  - class probabilities
- properties of probabilistic inference
  - chain rule of probability
  - marginalization
  - independence
  - Bayes rule

The chain rule of probability

- is an important consequence of the definition of conditional probability
  - note that, by recursive application of

$$P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y)$$

  - we can write

$$\boxed{\begin{aligned} P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = {} & P_{X_1|X_2,\ldots,X_n}(x_1|x_2,\ldots,x_n)\times \\ & \times P_{X_2|X_3\ldots,X_n}(x_2|x_3,\ldots,x_n)\times\ldots \\ & \times\ldots\times P_{X_{n-1}|X_n}(x_{n-1}|x_n)P_{X_n}(x_n) \end{aligned}}$$

- this is called the chain rule of probability
- it allows us to modularize inference problems

The chain rule of probability

- e.g. in the medical diagnosis scenario
  - what is the probability that you will be sick and have 104º of fever?

$$P_{Y,X_1}(sick,104) = P_{Y|X_1}(sick\,|\,104)P_{X_1}(104)$$

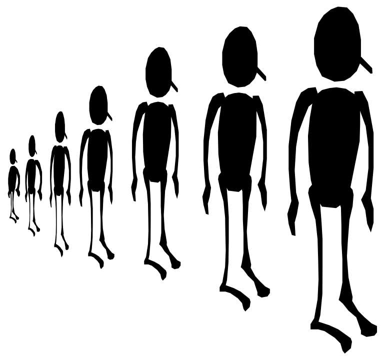
  - breaks down a hard question (prob of sick and 104) into two easier questions
  - Prob (sick|104): everyone knows that this is close to one

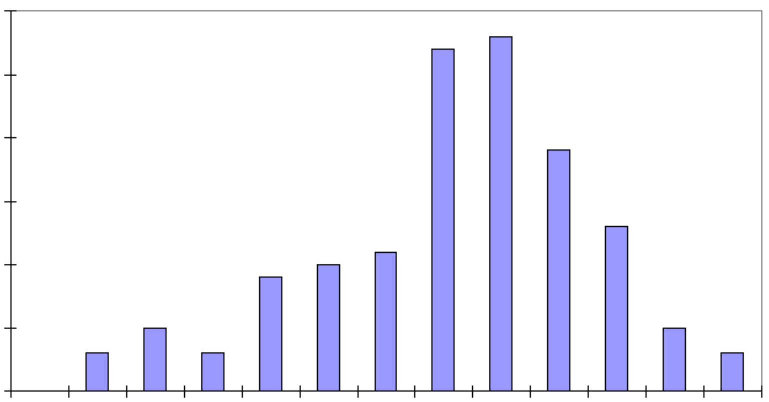
The chain rule of probability

- e.g. what is the probability that you will be sick and have 104º of fever?

$$P_{Y,X_1}(sick,104) = P_{Y|X_1}(sick\,|\,104)P_{X_1}(104)$$

  - Prob(104): still hard, but easier than P(sick,104) since we now only have one random variable (temperature)
  - does not depend on sickness, it is just the question "what is the probability that someone will have 104º?"
    - gather a number of people, measure their temperatures and make an histogram that everyone can use after that





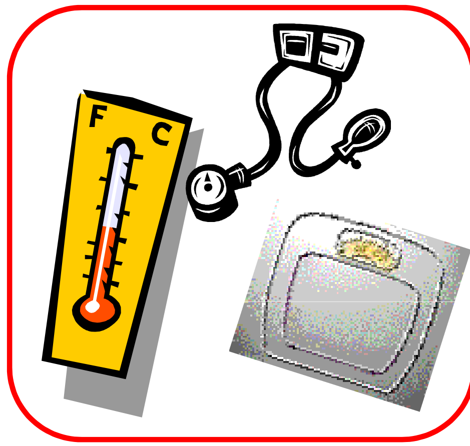
Tools for solving BDT problem

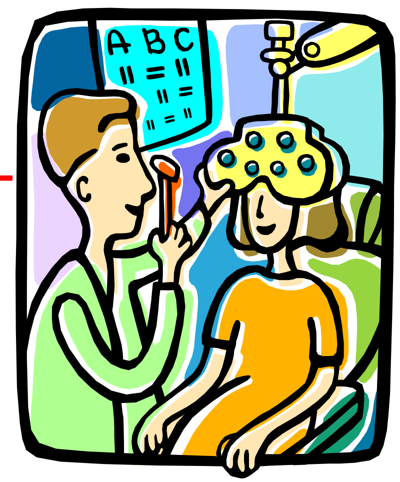
- probabilistic representations
  - joint distribution
  - class-conditional distributions
  - class probabilities

- properties of probabilistic inference
  - chain rule of probability
  - marginalization
  - independence
  - Bayes rule

Tools for solving BDT problems

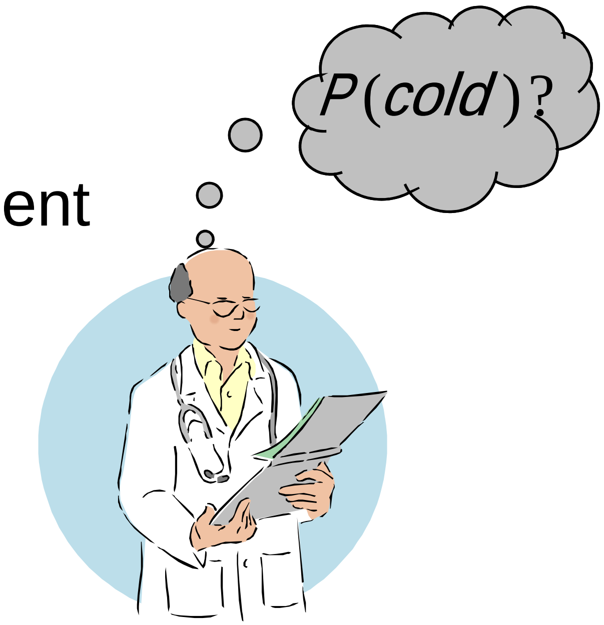
- frequently we have problems with multiple random variables
  - e.g. when in the doctor, you are mostly a collection of random variables
    - x_1: temperature
    - x_2: blood pressure
    - x_3: weight
    - x_4: cough





- we can summarize this as
  - a vector $\boldsymbol{X} = (x_1, \ldots, x_n)$ of n random variables
  - $P_{\boldsymbol{X}}(x_1, \ldots, x_n)$ is the joint probability distribution
- but frequently we only care about a subset of X

Marginalization



- what if I only want to know if the patient has a cold or not?
  - does not depend on blood pressure and weight
  - all that matters are fever and cough
  - that is, we need to know $P_{X1,X4}(a,b)$
- we marginalize with respect to a subset of variables
  - (in this case X_1 and X_4)
  - this is done by summing (or integrating) the others out

$$P_{X_1,X_4}(x_1,x_4) = \sum_{x_3,x_4} P_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)$$

$$P_{X_1,X_4}(x_1,x_4) = \int\int P_{X_1,X_2,X_3,X_4}(x_1,x_2,x_3,x_4)dx_2dx_3$$

Marginalization

- important equation:
  - seems trivial, but for large models is a major computational asset for probabilistic inference
  - for any question, there are lots of variables which are irrelevant
  - direct evaluation is frequently intractable
  - typically, we combine with the chain rule to explore independence relationships that will allow us to reduce computation
- independence:
  - X and Y are independent random variables if

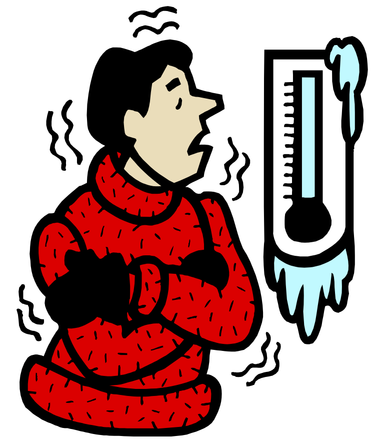
$$\boxed{P_{X|Y}(x \mid y) = P_X(x)}$$

Tools for solving BDT problem

- probabilistic representations
  - joint distribution
  - class-conditional distributions
  - class probabilities
- properties of probabilistic inference
  - chain rule of probability
  - marginalization
  - independence
  - Bayes rule

Independence

- very useful in the design of intelligent systems
  - frequently, knowing X makes Y independent of Z
  - e.g. consider the shivering symptom:
    - if you have temperature you sometimes shiver
    - it is a symptom of having a cold
    - but once you measure the temperature, the two become independent



$$
\begin{aligned}
P_{Y,X_1,S}(sick, 98, shiver) &= P_{Y|X_1,S}(sick \mid 98, shiver) \times \\
&\quad P_{S|X_1}(shiver \mid 98) P_{X_1}(98) \\
&= P_{Y|X_1}(sick \mid 98) \times \\
&\quad P_{S|X_1}(shiver \mid 98) P_{X_1}(98)
\end{aligned}
$$

- simplifies considerably the estimation of the probabilities

Independence

- combined with marginalization, enables efficient computation
  - e.g to compute P_Y(sick)
  - 1) marginalization

$$P_Y(sick) = \sum_s \int P_{Y,X_1,S}(sick, x, s)dx$$

  - 2) chain rule

$$P_Y(sick) = \sum_s \int P_{Y|X_1,S}(sick \mid x, s) P_{S|X_1}(s \mid x) P_{X_1}(x) dx$$

  - 3) independence

$$P_Y(sick) = \int P_{Y|X_1}(sick \mid x) P_{X_1}(x) \sum_s P_{S|X_1}(s \mid x) dx$$

  - dividing and grouping terms (divide and conquer) makes the integral simpler

Tools for solving BDT problem

- probabilistic representations
  - joint distribution
  - class-conditional distributions
  - class probabilities

- properties of probabilistic inference
  - chain rule of probability
  - marginalization
  - independence
  - Bayes rule

Tools for solving BDT problems

- Bayes rule

$$\boxed{P_{Y|X}(y|x) = \frac{P_{X|Y}(x|y)P_Y(y)}{P_X(x)}}$$

  - is the central equation of Bayesian inference
  - allows us to "switch" the relation between the variables
  - this is extremely useful
  - e.g. for medical diagnosis doctor needs to know

$$P_{Y|X}(disease\ y \mid symptom\ x)$$

  - this is very complicated because it is not causal
  - we are asking for the probability of cause given consequence

Tools for solving BDT problems

- Bayes rule transforms it into the probability of consequence given cause

$$P_{Y|X}(disease\ y\,|\,symptom\ x) =$$

$$= \frac{P_{X|Y}(symptom\ x\,|\,disease\ y)P_Y(disease\ y)}{P_X(symptom\ x)}$$

  and some other stuff

- note that $P_{X|Y}$(symptom x| disease y) is easy – you can get it out of any medical textbook
- what about the other stuff?
  - P_Y(disease y) does not depend on the patient – you can get it by collecting statistics over the entire population
  - P_X(symptom x) is a combination of the two (marginalization)

$$P_X(symptom\ x) = \sum_y P_{X|Y}(symptom\ x\,|\,disease\ y)P_Y(disease\ y)$$

Bayes rule

- Bayes rule allows us
  - to combine textbook knowledge with prior knowledge to compute the probability of cause given consequence
  - e.g. if you heard on the radio that there is an outbreak of "measles",
    - you increase the prior probability for the measles disease (cause)

$$P_Y(measles) \quad \uparrow\uparrow\uparrow$$

    - since (relation between cause and consequence)

$$P_{X|Y}(patient\ symptoms \mid measles)$$

does not change, Bayes rule will give you the "updated"

$$P_{Y|X}(measles \mid patient\ symptoms)$$

    - that accounts for the new information
    - this is hard if you work directly with the posterior probability

Tools for solving BDT problem

- probabilistic representations
  - joint distribution
  - class-conditional distributions
  - class probabilities
- properties of probabilistic inference
  - chain rule of probability
  - marginalization
  - independence
  - Bayes rule
- we are now ready to make optimal decisions!

Any questions?