Bayesian Decision Theory Tutorial

Tutorial 1 – the outline

- Bayesian decision making with discrete probabilities – an example
- Looking at continuous densities
- Bayesian decision making with continuous probabilities – an example
- The Bayesian Doctor Example

Example 1 – checking on a course

- A student needs to achieve a decision on which courses to take, based only on his first lecture.
- Define 3 categories of courses ω_j: **good, fair, bad**.
- From his previous experience, he knows:

Quality of the course	good	fair	bad
Probability (prior)	0.2	0.4	0.4

- These are **prior probabilities**.

Example 1 – continued

- The student also knows the **class-conditionals**:

$\Pr(x\|\omega_j)$	good	fair	bad
Interesting lecture	0.8	0.5	0.1
Boring lecture	0.2	0.5	0.9

- The **loss function** is given by the matrix

$\lambda(a_i\|\omega_j)$	good course	fair course	bad course
Taking the course	0	5	10
Not taking the course	20	5	0

Example 1 – continued

- The student wants to make an optimal decision=> minimal possible $R(\alpha)$,

  while α: x->{***take the course***, ***drop the course***}

- The student needs to minimize the conditional risk;

$$R(\alpha_i \mid x)=\sum_{j=1}^{c}\lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x)$$

given

given

compute $$P(\omega_j \mid x)=\frac{\boxed{P(x \mid \omega_j)}\boxed{P(\omega_j)}}{\boxed{P(x)}}$$

compute

Example 1 : compute P(x)

- The probability to get an “interesting lecture” (x= interesting):

  Pr(interesting)= Pr(interesting|good course)* Pr(good course)

  + Pr(interesting|fair course)* Pr(fair course)

  + Pr(interesting|bad course)* Pr(bad course)

  =0.8*0.2+0.5*0.4+0.1*0.4=0.4

- Consequently, Pr(boring)=1-0.4=0.6

Example 1 : compute $P(\omega_j \mid x)$

Suppose the lecture was interesting. Then we want to compute the **posterior** probabilities of each one of the 3 possible "states of nature".

$$\Pr(\text{good course}|\text{interesting lecture})$$
$$= \frac{\Pr(\text{interesting}|\text{good})\Pr(\text{good})}{\Pr(\text{interesting})} = \frac{0.8*0.2}{0.4} = 0.4$$

$$\Pr(\text{fair}|\text{interesting})$$
$$= \frac{\Pr(\text{interesting}|\text{fair})\Pr(\text{fair})}{\Pr(\text{interesting})} = \frac{0.5*0.4}{0.4} = 0.5$$

- We can get Pr(bad|interesting)=0.1 either by the same method, or by noting that it complements to 1 the above two.

Example 1 $R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$

- The student needs to minimize the conditional risk;

  take the course:

  R(taking| interesting)= λ(taking| good) Pr(good| interesting)
  + λ(taking| fair) Pr(fair| interesting)
  + λ(taking| bad) Pr(bad| interesting)
  = 0.4*0+0.5*5+0.1*10=3.5

  or drop it:

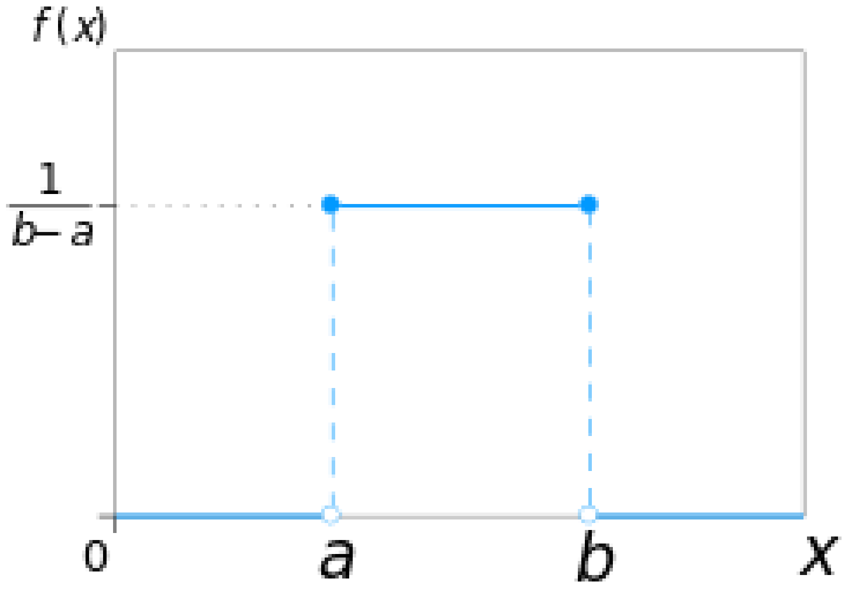
  R(droping| interesting)= λ(droping| good) Pr(good| interesting)
  + λ(droping| fair) Pr(fair| interesting)
  + λ(droping| bad) Pr(bad| interesting)
  = 0.4*20+0.5*5+0.1*0=10.5

Constructing an optimal decision function

- So, if the first lecture was interesting, the student will minimize the conditional risk <u>by taking the course</u>.
- In order to construct the full decision function, we need to define the risk minimization action for the case of boring lecture, as well.

Do it!

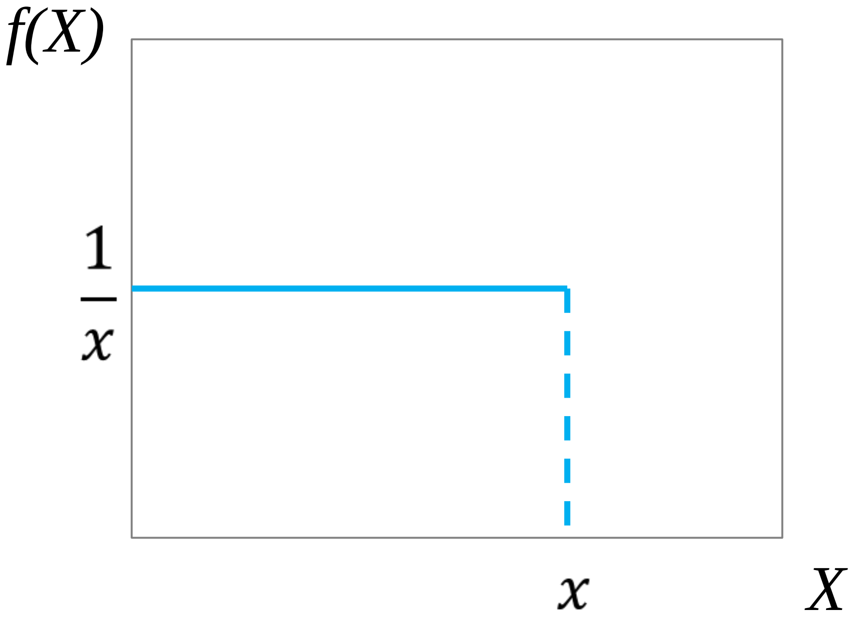
Uniform Distribution



Example 2 – continuous density

- Let X be a real value r.v., representing a number randomly picked from the interval [0,1]; its distribution is known to be uniform.
- Then let Y be a real r.v. whose value is chosen at random from $[0, X]$ also with uniform distribution.
- We are presented with the value of Y, and need to "guess" the most "likely" value of X.
- In a more formal fashion: given the value of Y, find the probability density function of X and determine its maxima.

Uniform Distribution



Example 2 – continued

- What we look for is P(X=x | Y=y) – that is, the **p.d.f**.
- The class-conditional (given the value of X):

$$P(Y = y \mid X = x) = \begin{cases} \dfrac{1}{x} & y \le x \le 1 \\ 0 & y > x \end{cases}$$

- For the given evidence:

(using total probability)

$$P(Y = y) = \int_y^1 \frac{1}{x} dx = \ln\left(\frac{1}{y}\right)$$

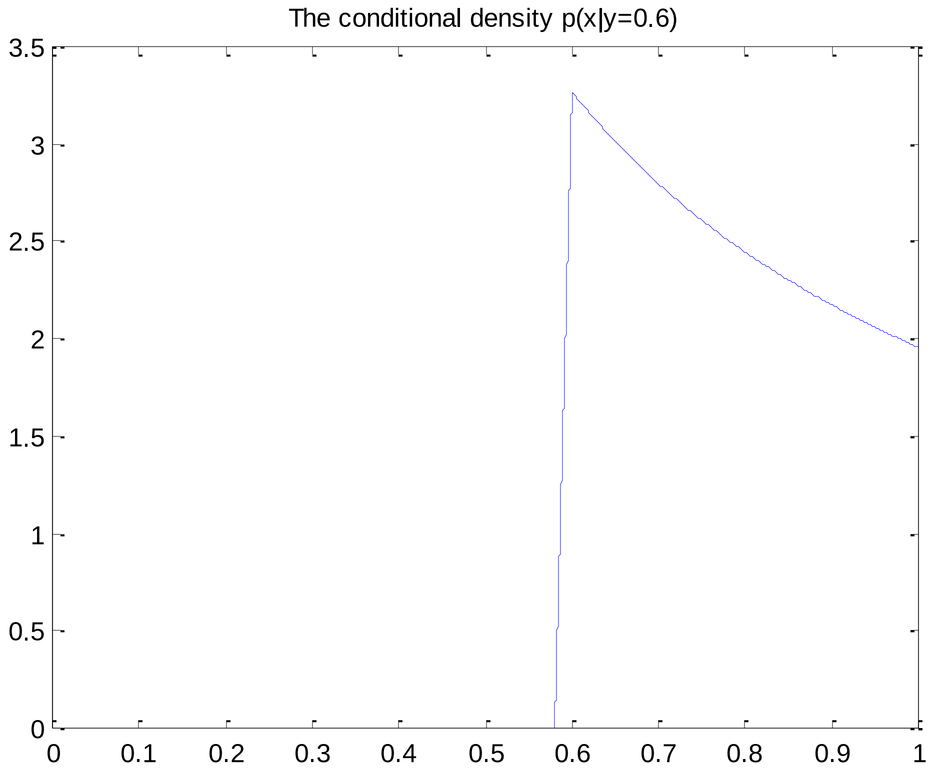
Example 2 – conclusion

- Applying Bayes' rule:

$$p(x \mid y) = \frac{p(y \mid x)\, p(x)}{p(y)} = \frac{\frac{1}{x} 1}{\ln\left(\frac{1}{y}\right)}$$

- This is monotonically decreasing function, over [y,1].
- So (informally) the most "likely" value of X (the one with highest probability density value) is *X*=*y*.

Illustration – conditional p.d.f.



Example 3: hiring a secretary

- A manager needs to hire a new secretary, and a good one.
- Unfortunately, good secretary are hard to find:

$$\Pr(w_g)=0.2, \quad \Pr(w_b)=0.8$$

- The manager decides to use a new test. The grade is a real number in the range from 0 to 100.
- The manager's estimation of the possible losses:

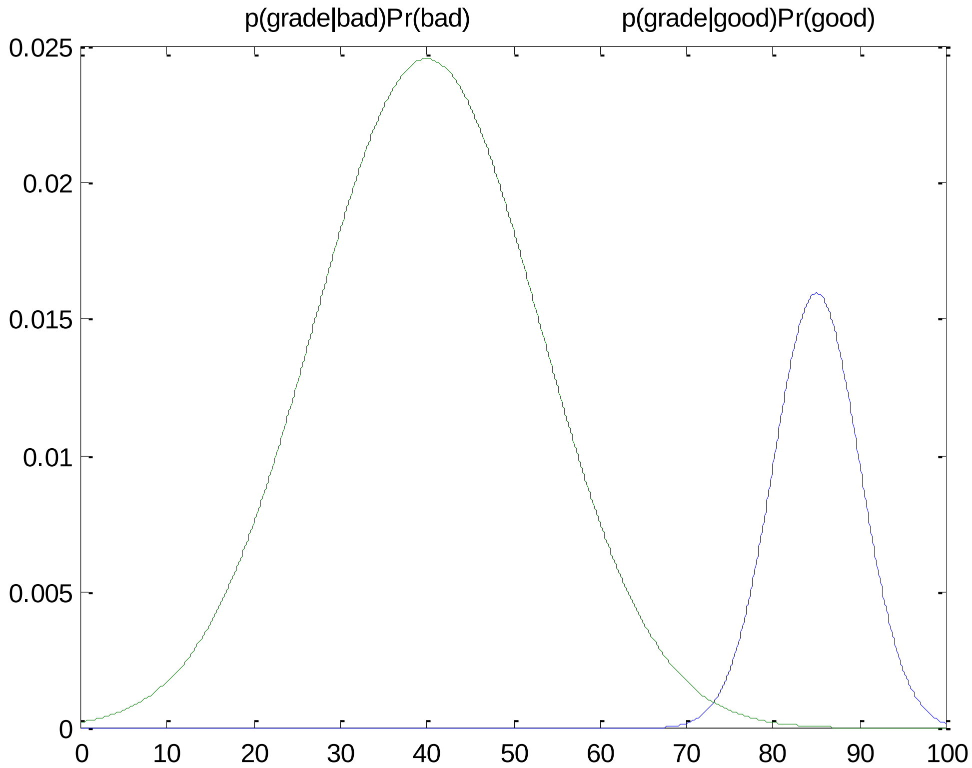
$\lambda(decision, w_i)$	w_g	w_b
Hire	0	20
Reject	5	0

Example 3: continued

- The class conditional densities are known to be approximated by a normal p.d.f.:

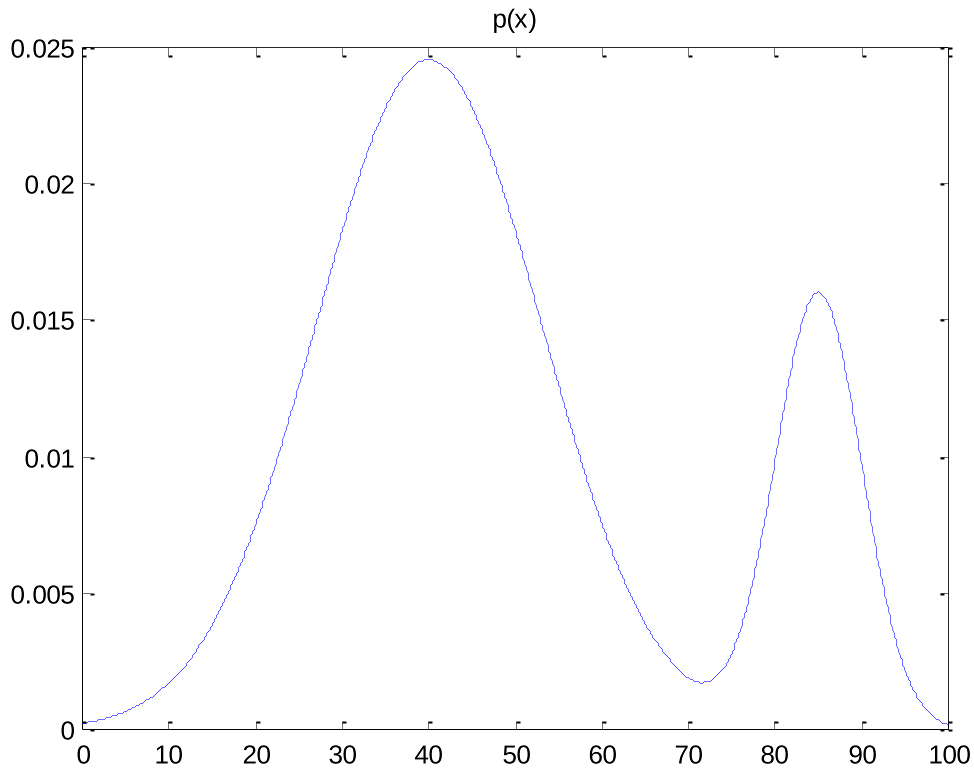
$$p(grade \mid good\ secretary) \sim N(85,5)$$

$$p(grade \mid bad\ secretary) \sim N(40,13)$$



Example 3: continued

- The resulting probability density for the grade looks as follows: $p(x)=p(\,x|w_b\,)p(\,w_b\,)+p(\,x|w_g\,)p(\,w_g\,)$



Example 3: continued

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

- We need to know for which grade values hiring the secretary would minimize the risk:

$$R(\text{hire} \mid x) < R(\text{reject} \mid x) \Leftrightarrow$$

$$p(w_b \mid x)\lambda(\text{hire}, w_b) + p(w_g \mid x)\lambda(\text{hire}, w_g)$$

$$< p(w_b \mid x)\lambda(\text{reject}, w_b) + p(w_g \mid x)\lambda(\text{reject}, w_g) \Leftrightarrow$$

$$[\lambda(\text{hire}, w_b) - \lambda(\text{reject}, w_b)] \cdot p(w_b \mid x) < [\lambda(\text{reject}, w_g) - \lambda(\text{hire}, w_g)] p(w_g \mid x)$$

- The posteriors are given by

$$p(w_i \mid x) = \frac{p(x \mid w_i) p(w_i)}{p(x)}$$

Example 3: continued

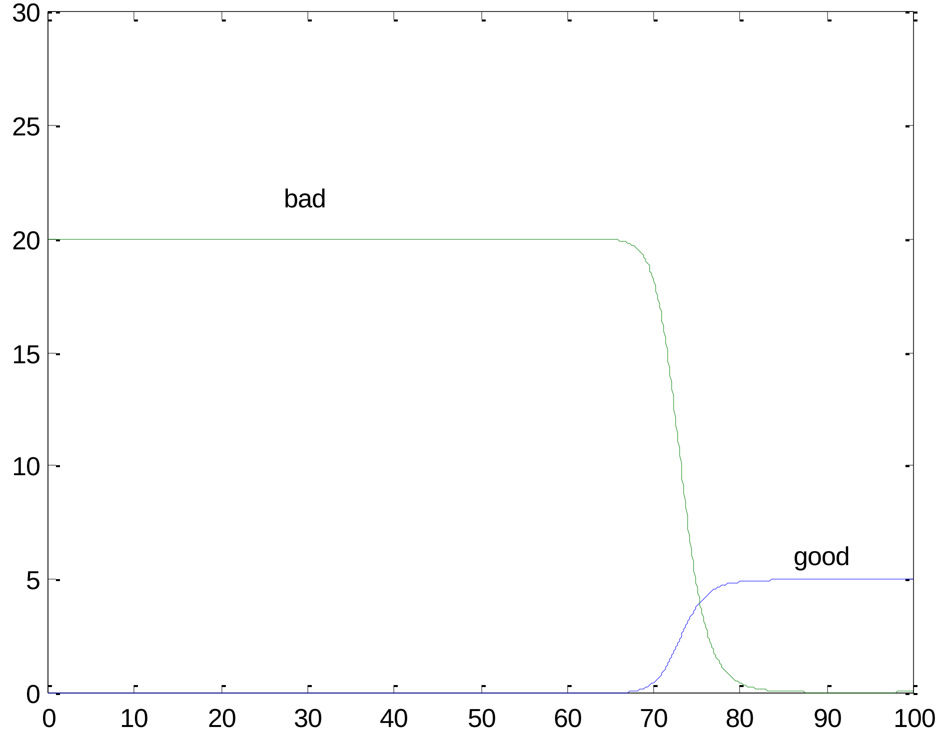
- The posteriors scaled by the loss differences,

$$[\lambda(\text{hire}, w_b) - \lambda(\text{reject}, w_b)] \cdot p(w_b \mid x)$$

and

$$[\lambda(\text{reject}, w_g) - \lambda(\text{hire}, w_g)] \cdot p(w_g \mid x)$$

look like:



Example 3: continued

- Numerically, we have:

$$p(x) = \frac{0.2}{5\sqrt{2\pi}} e^{-\frac{(x-85)^2}{2\cdot 5^2}} + \frac{0.8}{13\sqrt{2\pi}} e^{-\frac{(x-40)^2}{2\cdot 13^2}}$$

$$p(w_b \mid x) = \frac{\frac{0.8}{13\sqrt{2\pi}} e^{-\frac{(x-40)^2}{2\cdot 13^2}}}{p(x)}, \qquad p(w_g \mid x) = \frac{\frac{0.2}{5\sqrt{2\pi}} e^{-\frac{(x-85)^2}{2\cdot 5^2}}}{p(x)}$$

- We need to solve $20\, p(w_b \mid x) > 5\, p(w_g \mid x)$

- Solving numerically yields one solution in [0, 100]:

$$x \approx 76$$

The Bayesian Doctor Example

A person doesn't feel well and goes to a doctor.

Assume two states of nature:

ω_1 : The person has a common flue.

ω_2 : The person is really sick (a vicious bacterial infection).

The doctor's ***prior*** is: $p(\omega_1) = 0.9$ $p(\omega_2) = 0.1$

This doctor has two possible actions: "prescribe" hot tea or antibiotics. Doctor can use prior and predict optimally: always flue. Therefore doctor will always prescribe hot tea.

The Bayesian Doctor - Cntd.

- But there is very high risk: Although this doctor can diagnose with very high rate of success using the prior, (s)he can lose a patient once in a while.
- Denote the two possible actions:

  a_1 = prescribe hot tea

  a_2 = prescribe antibiotics
- Now assume the following cost (loss) matrix:

$$\lambda_{i,j} = \begin{array}{c|c|c|} & \overset{\text{flue}}{\omega_1} & \overset{\text{bacteria}}{\omega_2} \\ \hline a_1 & 0 & 10 \\ \hline a_2 & 1 & 0 \\ \hline \end{array}$$

The Bayesian Doctor - Cntd.

- Choosing α_1 results in ***expected risk*** of

$$R(a_1) = p(\omega_1)\cdot\lambda_{1,1} + p(\omega_2)\cdot\lambda_{1,2}$$

$$= 0 + 0.1\cdot 10 = 1$$

- Choosing α_2 results in expected risk of

$$R(a_2) = p(\omega_1)\cdot\lambda_{2,1} + p(\omega_2)\cdot\lambda_{2,2}$$

$$= 0.9\cdot 1 + 0 = 0.9$$

- So, considering the costs it's much better (and optimal!) to always give antibiotics.

The Bayesian Doctor - Cntd.

- But doctors can do more. For example, they can take some ***observations.***
- A reasonable observation is to perform a blood test.
- Suppose the possible results of the blood test are:
  x_1 = negative (no bacterial infection)
  x_2 = positive (infection)
- But blood tests can often fail. Suppose
  (***class conditional*** probabilities.)

infection $p(x_1 \mid \omega_2) = 0.3$ $p(x_2 \mid \omega_2) = 0.7$

flue $p(x_2 \mid \omega_1) = 0.2$ $p(x_1 \mid \omega_1) = 0.8$

The Bayesian Doctor - Cntd.

- Define the conditional risk given the observation

$$R(a_i \mid x) = \sum_{\omega_j} p(\omega_j \mid x) \cdot \lambda_{i,j}$$

- We would like to compute the conditional risk for each action and observation so that the doctor can choose an optimal action that minimizes risk.
- How can we compute $p(\omega_j \mid x)$?
- We use the class conditional probabilities and Bayes inversion rule.

The Bayesian Doctor - Cntd.

- Let's calculate first p(x_1) and p(x_2)

$$\mathrm{p}(\mathrm{x}_1) = \mathrm{p}(\mathrm{x}_1 \mid \omega_1) \cdot \mathrm{p}(\omega_1) + \mathrm{p}(\mathrm{x}_1 \mid \omega_2) \cdot \mathrm{p}(\omega_2)$$
$$= 0.8 \cdot 0.9 + 0.3 \cdot 0.1$$
$$= 0.75$$

- p(x_2) is complementary to p(x_1) , so $\mathrm{p}(\mathrm{x}_2) = 0.25$

The Bayesian Doctor - Cntd.

$$R(\alpha_1 \mid x_1) = p(\omega_1 \mid x_1) \cdot \lambda_{1,1} + p(\omega_2 \mid x_1) \cdot \lambda_{1,2}$$

$$= 0 + p(\omega_2 \mid x_1) \cdot 10$$

$$= 10 \cdot \frac{p(x_1 \mid \omega_2) \cdot p(\omega_2)}{p(x_1)}$$

$$= 10 \cdot \frac{0.3 \cdot 0.1}{0.75} = 0.4$$

$$R(\alpha_2 \mid x_1) = p(\omega_1 \mid x_1) \cdot \lambda_{2,1} + p(\omega_2 \mid x_1) \cdot \lambda_{2,2}$$

$$= p(\omega_1 \mid x_1) \cdot 1 + p(\omega_2 \mid x_1) \cdot 0$$

$$= \frac{p(x_1 \mid \omega_1) \cdot p(\omega_1)}{p(x_1)}$$

$$= \frac{0.8 \cdot 0.9}{0.75} = 0.96$$

The Bayesian Doctor - Cntd.

$$R(\alpha_1 \mid x_2) = p(\omega_1 \mid x_2) \cdot \lambda_{1,1} + p(\omega_2 \mid x_2) \cdot \lambda_{1,2}$$

$$= 0 + p(\omega_2 \mid x_2) \cdot 10$$

$$= 10 \cdot \frac{p(x_2 \mid \omega_2) \cdot p(\omega_2)}{p(x_2)}$$

$$= 10 \cdot \frac{0.7 \cdot 0.1}{0.25} = 2.8$$

$$R(\alpha_2 \mid x_2) = p(\omega_1 \mid x_2) \cdot \lambda_{2,1} + p(\omega_2 \mid x_2) \cdot \lambda_{2,2}$$

$$= p(\omega_1 \mid x_2) \cdot 1 + p(\omega_2 \mid x_2) \cdot 0$$

$$= \frac{p(x_2 \mid \omega_1) \cdot p(\omega_1)}{p(x_2)}$$

$$= \frac{0.2 \cdot 0.9}{0.25} = 0.72$$

The Bayesian Doctor - Cntd.

- To summarize:

$$R(\alpha_1 \mid x_1) = 0.4$$
$$R(\alpha_2 \mid x_1) = 0.96$$
$$R(\alpha_1 \mid x_2) = 2.8$$
$$R(\alpha_2 \mid x_2) = 0.72$$

- Whenever we encounter an observation x, we can minimize the expected loss by minimizing the conditional risk.
- Makes sense: Doctor chooses hot tea if blood test is negative, and antibiotics otherwise.

Optimal Bayes Decision Strategies

- A ***strategy*** or ***decision function*** $\alpha(x)$ is a mapping from observations to actions.
- The ***total risk*** of a decision function is given by

$$E_{p(x)}[R(\alpha(x)\mid x)] = \sum_{x} p(x)\cdot R(\alpha(x)\mid x)$$

- A decision function is ***optimal*** if it minimizes the total risk. This optimal total risk is called ***Bayes risk***.

- In the Bayesian doctor example:
  – Total risk if doctor always gives antibiotics(a_2): 0.9
  – Bayes risk: 0.48   How have we got it?