Naïve Bayes

Decision theory

- Decision theory is the study of making decisions that have a significant impact

- Decision-making is distinguished into:
  - Decision-making under certainty
  - Decision-making under non-certainty
    - Decision-making under risk
    - Decision-making under uncertainty

Probability theory

- Most decisions have to be taken in the presence of uncertainty
- Basic elements of probability theory:
  - Random variables describe aspects of the world whose state is initially unknown
  - Each random variable has a domain of values that it can take on (discrete, boolean, continuous)
  - An atomic event is a complete specification of the state of the world

Probability Theory

- All probabilities are between 0 and 1
- The sum of probabilities for the atomic events of a probability space must sum up to 1

Prior

- **Priori Probabilities** or Prior reflects our prior knowledge of how likely an event occurs.

Class Conditional probability (posterior)

- When we have information concerning previously unknown random variables then we use **posterior** or conditional probabilities: P(a|b) the **probability of a given event a that we know b**

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$$

- Alternatively this can be written (the product rule):

$$P(a \wedge b) = P(a|b)P(b)$$

Bayes rule

- The product rule can be written as:
- $P(a \wedge b)=P(a|b)P(b)$
- $P(a \wedge b)=P(b|a)P(a)$

- By equating the right-hand sides:

$$P(b \mid a)=\frac{P(a \mid b)P(b)}{P(a)}$$

Posterior Probabilities

- Define p(cj/x) as the posteriori probability
- We use Baye’s formula to convert the prior to posterior probability

$$p(cj \mid x) = \frac{p(x \mid cj)\, p(cj)}{p(x)}$$

Bayes Classifiers

- Bayesian classifiers use **Bayes theorem**, which says

$$p(cj \mid x) = \frac{p(x \mid cj)\, p(cj)}{p(x)}$$

- ***p*(*cj* | *x*)** = probability of instance x being in class *cj*,

  This is what we are trying to compute

- ***p*(x | c*j*)** = probability of generating instance *x* given class *cj*,

  We can imagine that being in class c*j* causes you to have feature x with some probability

- ***p(cj)*** = probability of occurrence of class *cj*,

  This is just how frequent the class c*j* is in our database

Bayes Formula

- Suppose the priors P(cj) and conditional densities p(x|cj) are known,

$$\underbrace{P(C \mid \mathbf{x})}_{\textit{posterior}} = \frac{\overbrace{P(C)}^{\textit{prior}}\,\overbrace{p(\mathbf{x} \mid C)}^{\textit{likelihood}}}{\underbrace{p(\mathbf{x})}_{\textit{evidence}}}$$

Bayesian Decision Theory

Tradeoffs between various decisions using probabilities and costs that accompany such decisions.

Example: **Patient has trouble breathing**

- **Decision**: Asthma versus Lung cancer

1. Decide lung cancer when person has asthma
   - Cost: moderately high (e.g., order unnecessary tests, scare patient)
2. Decide asthma when person has lung cancer
   - **Cost: very high (e.g., lose opportunity to treat cancer at early stage, death)**

Decision Rules

Progression of decision rules:

1. Decide based on prior probabilities
2. Decide based on posterior probabilities
3. Decide based on risk

Fish Sorting Example

- C ➔ class

C=c1 (sea bass)

C=c2 (salmon)

- P(c1) is the prior probability that the next fish is a sea bass
- P(c2) is the prior probability that the next fish is a salmon

Decision based on prior probabilities

- Assume $P(c1) + P(c2) = 1$
- Decision ??
- Decide ➔
  - C1 if $P(c1) > P(c2)$
  - C2 otherwise
- Error probability
  $p(error) = \min(P(c1), P(c2))$

Decision based on class conditional probabilities

- Let x be a continuous random variable
- Define p(x/cj) as the conditional probability density (j=1,2)
- P(x/c1) and P(x/c2) describe the difference in measurement between populations of sea bass and Solomon

Making a Decision

- Decision ??? (After observing x value)
- Decide :
  - C1 if $P(c1/x) > P(c2/x)$
  - C2 otherwise
- $P(c1/x) + P(c2/x)=1$

Probability of Error

- $P(error/x)$:
  - $P(c1/x)$ if we decide c2
  - $P(c2/x)$ if we decide c1
- $P(error/x) = \min \{P(c1/x), P(c2/x)\}$

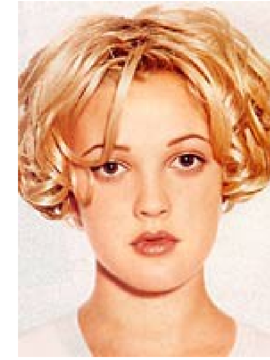
Assume that we have two classes

$c1$ = male, and $c2$ = female.

We have a person whose sex we do not know, say "*drew*" or d.

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, I.e which is greater $p(\text{male} \mid drew)$ or $p(\text{female} \mid drew)$

(Note: "Drew can be a male or female name")



Drew Barrymore

Drew Carey

What is the probability of being called "*drew*" given that you are a male?

What is the probability of being a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

What is the probability of being named "*drew*"? (actually irrelevant, since it is same for all classes)

Officer Drew

This is Officer Drew (who arrested abc in 2007). Is Officer Drew a Male or Female?

Luckily, we have a small database with names and sex.

We can use it to apply Bayes rule…

HERE ONLY MULTIPLY

HERE WE MULTIPLY ONLY BECAUSE LIKLYHOOD IS CALCULATED BUT BELOW WE SHOULD WRITE PRIOR PROBABILTY, HERE COMPARISON BASED ON CONDITIONAL PROBABILTY

$$p(cj \mid d) = \frac{p(d \mid cj)\, p(cj)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Officer Drew

$$p(cj \mid d) = \frac{p(d \mid cj)\, p(cj)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid \textit{drew}) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid \textit{drew}) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female.

Officer Drew

Officer Drew IS a female!

$$p(\text{male} \mid drew) = \frac{1/3 \;*\; 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 \;*\; 5/8}{3/8} = \frac{0.250}{3/8}$$

Generalized Bayesian Decision Theory

- More than one observation x
  - Replace scalar x with vector **x**
- Allowing actions other than just decision?
  - Allow the possibility of rejection
- Different risks in the decision
  - Define how costly each action is

Bayesian Decision Theory

- Let {c1,c2,,..cn} be classes/states
- Let {α1, α2, α3,.. ,αa} be finite set of a possible actions.
- Let λ(αi/cj) be the loss incurred for taking action αi when the class is cj.
- Let **x** be random variable (vector).

Conditional Risk

- Suppose we observe **x** and take action αi.
- If the true class is cj, we incur the loss λ(αi/cj).
- The expected loss (conditional risk) with taking action αi is

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

STUDY TOTAL PROBABILITY CONCEPT ???

Minimum-Risk Classification

- For every **x the** decision function α(**x**) assumes one of the **a values α1, ..., αa.**
- **The overall risk R is the** expected loss associated with a given decision rule.
- The general decision rule **α(x)** tells us which action to take for **x.**
- We want to find out the decision rule that minimizes the overall risk

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}.$$

- Bayes decision rule, minimizes the overall risk by selecting the action **αi for which R(αi/x) is minimum.**

Two-category classification

$\alpha 1$: deciding $c1$

$\alpha 2$: deciding $c2$

$\lambda ij = \lambda(\alpha i \mid cj)$

loss incurred for deciding ci when the true state of nature is cj

Conditional risk:

$$R(\alpha 1 \mid x) = \lambda 11 P(c1 \mid x) + \lambda 12 P(c2 \mid x)$$

$$R(\alpha 2 \mid x) = \lambda 21 P(c1 \mid x) + \lambda 22 P(c2 \mid x)$$

The rule is the following:

if $R(\alpha 1 \mid x) < R(\alpha 2 \mid x)$

action $\alpha 1$: "decide c*1*" is take

PUT VALUE IN THIS INEQUALITY
AND THEN SUBTRACT

This results in the equivalent rule :

Decide c*1* if:

$$(\lambda 21-\lambda 11\ )\ P(c1 \mid x) > (\lambda 12-\lambda 22\ )\ P(c2 \mid x)$$

By Bayes formula

$$(\lambda 21- \lambda 11)\ \mathbf{P(x \mid c1)\ P(c1)} > (\lambda 12- \lambda 22)\ \mathbf{P(x \mid c2)\ P(c2)}$$

Likelihood ratio

$$if\ \frac{P(x \mid c_1)}{P(x \mid c_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(c_2)}{P(c_1)}$$

Then take action $\alpha 1$ (decide *c1*)

Otherwise take action $\alpha 2$ (decide *c2*)

Example

- Suppose selection of c1 and c2 has same probability:

  P(c1)=p(c2)=1/2

Assume that the lo $L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$ he form

- If misclassification of patterns that come from c2 is considered to have serious

- Thus, patterns are assigned to c2 class if

$$P(x \mid c_2) > \frac{\lambda_{21}}{\lambda_{12}} P(x \mid c_1)$$

- That is, *P(x / c1)* ***is multiplied by*** a factor less than 1

The Bayesian Doctor Example

A person doesn't feel well and goes to the 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 doctors ***prior*** is: $p(\omega_1) = 0.9$ $p(\omega_2) = 0.1$

W1➔c1 and w2➔c2

The Bayesian Doctor Example

A person doesn't feel well and goes to the 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 doctors ***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.

W1 ➔ c1 and w2 ➔ c2

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:

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

The Bayesian Doctor - Cntd.

- Choosing a_1 results in ***expected risk*** of

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

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

The Bayesian Doctor - Cntd.

- Choosing a_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 a_2 results in expected risk of

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

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

The Bayesian Doctor - Cntd.

- Choosing a_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 a_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:

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)

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)

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

The Bayesian Doctor - Cntd.

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

$$\begin{aligned} p(x_1) &= p(x_1 \mid \omega_1) \cdot p(\omega_1) + p(x_1 \mid \omega_2) \cdot p(\omega_2) \\ &= 0.8 \cdot 0.9 + 0.3 \cdot 0.1 \\ &= 0.75 \end{aligned}$$

- p(x_2) is complementary to p(x_1) , so $p(x_2) = 0.25$

The Bayesian Doctor - Cntd.

$$\begin{aligned}
\mathrm{R}(a_1 \mid x_1) &= p(\omega_1 \mid x_1) \cdot \lambda_{1,1} + p(\omega_2 \mid x_1) \cdot \lambda_{1,2} \\
&= 0 + p(\omega_2 \mid x_1) \cdot 10 \\
&= 10 \cdot \frac{p(x_1 \mid \omega_2) \cdot p(\omega_2)}{p(x_1)} \\
&= 10 \cdot \frac{0.3 \cdot 0.1}{0.75} = 0.4
\end{aligned}$$

The Bayesian Doctor - Cntd.

$$\begin{aligned}
\mathrm{R}(a_1 \mid x_1) &= p(\omega_1 \mid x_1)\cdot\lambda_{1,1} + p(\omega_2 \mid x_1)\cdot\lambda_{1,2} \\
&= 0 + p(\omega_2 \mid x_1)\cdot 10 \\
&= 10\cdot\frac{p(x_1 \mid \omega_2)\cdot p(\omega_2)}{p(x_1)} \\
&= 10\cdot\frac{0.3\cdot 0.1}{0.75} = 0.4
\end{aligned}$$

$$\begin{aligned}
R(a_2 \mid x_1) &= p(\omega_1 \mid x_1)\cdot\lambda_{2,1} + p(\omega_2 \mid x_1)\cdot\lambda_{2,2} \\
&= p(\omega_1 \mid x_1)\cdot 1 + p(\omega_2 \mid x_1)\cdot 0 \\
&= \frac{p(x_1 \mid \omega_1)\cdot p(\omega_1)}{p(x_1)} \\
&= \frac{0.8\cdot 0.9}{0.75} = 0.96
\end{aligned}$$

The Bayesian Doctor - Cntd.

$$
\begin{aligned}
R(a_1 \mid x_2) &= p(\omega_1 \mid x_2) \cdot \lambda_{1,1} + p(\omega_2 \mid x_2) \cdot \lambda_{1,2} \\
&= 0 + p(\omega_2 \mid x_2) \cdot 10 \\
&= 10 \cdot \frac{p(x_2 \mid \omega_2) \cdot p(\omega_2)}{p(x_2)} \\
&= 10 \cdot \frac{0.7 \cdot 0.1}{0.25} = 2.8
\end{aligned}
$$

The Bayesian Doctor - Cntd.

$$R(a_1 \mid x_2) = p(\omega_1 \mid x_2) \cdot \lambda_{1,1} + p(\omega_2 \mid x_2) \cdot \lambda_{1,2}$$
$$= 0 + p(\omega_2 \mid x_2) \cdot 10$$
$$= 10 \cdot \frac{p(x_2 \mid \omega_2) \cdot p(\omega_2)}{p(x_2)}$$
$$= 10 \cdot \frac{0.7 \cdot 0.1}{0.25} = 2.8$$

$$R(a_2 \mid x_2) = p(\omega_1 \mid x_2) \cdot \lambda_{2,1} + p(\omega_2 \mid x_2) \cdot \lambda_{2,2}$$
$$= p(\omega_1 \mid x_2) \cdot 1 + p(\omega_2 \mid x_2) \cdot 0$$
$$= \frac{p(x_2 \mid \omega_1) \cdot p(\omega_1)}{p(x_2)}$$
$$= \frac{0.2 \cdot 0.9}{0.25} = 0.72$$

The Bayesian Doctor - Cntd.

- To summarize:

$$R(a_1 \mid x_1) = 0.4$$

$$R(a_2 \mid x_1) = 0.96$$

$$R(a_1 \mid x_2) = 2.8$$

$$R(a_2 \mid x_2) = 0.72$$

The Bayesian Doctor - Cntd.

- To summarize:

$$R(a_1 \mid x_1) = 0.4$$
$$R(a_2 \mid x_1) = 0.96$$
$$R(a_1 \mid x_2) = 2.8$$
$$R(a_2 \mid x_2) = 0.72$$

- Whenever we encounter an observation x, we can minimize the expected loss by minimizing the conditional risk.

The Bayesian Doctor - Cntd.

- To summarize:

$$R(a_1 \mid x_1) = 0.4$$
$$R(a_2 \mid x_1) = 0.96$$
$$R(a_1 \mid x_2) = 2.8$$
$$R(a_2 \mid x_2) = 0.72$$

- 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.

Advantages/Disadvantages of Naïve Bayes

- Advantages
  - Fast to train (single scan). Fast to classify
  - Handles real and discrete data
  - Handles streaming data well
- Disadvantages
  - Assumes independence of features