Bayesian Decision Theory

Sanne Nygaard

Outline

- Modelling
- Joint probabilities and Bayes' Rule
- Bayesian classification
- Losses, Risk, and Rejection
- Association rules

Modelling

We model data as random process to make up for noise and missing information

How would we model a coin toss?

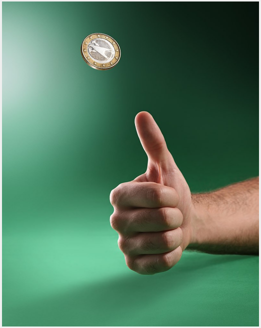


Image credit: ICMA Photos

Example: Coin Toss

- There is the *observable variable*, in this case the outcome: $\mathit{Result} \in \{\mathit{Heads}, \mathit{Tails}\}$

Example: Coin Toss

- There is the *observable variable*, in this case the outcome: $\mathit{Result} \in \{\mathit{Heads}, \mathit{Tails}\}$
- There are *unobservable variables* that we don't have access to, e.g.:
  - Coin composition
  - Precise position and orientation
  - Strength and direction of forces
- If we had access to all info we might calculate the outcome

Example: Roulette

Image credit: Ralf Roletschek

Example: Roulette

Image credit: A rather dodgy website

Example: Coin Toss

- Modelled using the Bernoulli distribution:
  Result $\in \{Heads, Not\ Heads\} \ \rightarrow$
  Bernoulli random variable: $X \in \{1, 0\}$

$$P\{X = x\} = p^x(1 - p)^{(1-x)}$$

p = probability of '1'
1 - p = probability of '0'

Estimation

What if we don't know p? How do we find it?

Example: Coin Toss

We get a sample of observed outcomes:

Sample: $\boldsymbol{X} = \{x^t\}_{t=1}^N$
Estimation:

$$p = \frac{\#\{Heads\}}{\#\{Tosses\}} = \sum_t \frac{x^t}{N}$$

This illustrates the frequentist approach to probability:

- Probabilities as frequencies of outcomes from repeated experiments
- There is a true value for the parameter p

Example: Coin Toss

Estimation example:

$$\boldsymbol{X} = \{1, 1, 1, 0, 0, 1, 1, 0, 1\}$$
$$\hat{p} = \frac{\#\{Heads\}}{\#\{Tosses\}} = \frac{6}{9}$$

Image source: wikipedia.org

If instead of nine observations:

$$\boldsymbol{X} = \{1, 1, 1, 0, 0, 1, 1, 0, 1\}$$
$$\hat{p} = \frac{6}{9}$$

We only had the first three:

$$\boldsymbol{X} = \{1, 1, 1\}$$
$$\hat{p} = \frac{3}{3} = 1$$

Example: Coin Toss

Predictions

Minimize errors, choose highest probability outcome

Prediction: Heads if $p > \frac{1}{2}$, Tails otherwise
Probability of error: $1 - P(\textit{our choice})$

Example: Coin Toss

In summary:

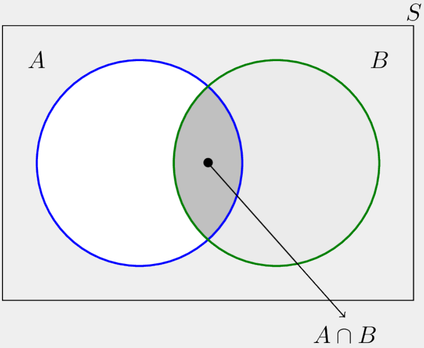
1. Choose a model
2. Use observed data to estimate model parameters
3. Use the trained model to make predictions

Intersection: $P(A \cap B)$

$$P(A \cap B) = P(A|B)P(B)$$
$$= P(B|A)P(A)$$

For independent variables:

$$P(A|B) = P(A)$$
$$P(A, B) = P(A)P(B)$$



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Image from: www.probabilitycourse.com

This gives us *Bayes' rule*:

$$P(B|A) = \frac{p(A|B)P(B)}{P(A)}$$

Example: *The Decaying Fruit Basket*

- A not-so-fresh fruit basket with N = 25 pieces of fruits: Oranges, Apples, Peaches
- Some of them are good, some have gone bad
- The numbers are:

	Oranges	Apples	Peaches
Good	4	7	3
Bad	3	1	7

- Joint distribution over two random variables:
- Fruit type: $X \in \{Oranges, Apples, Peaches\}$
- Freshness: $Y \in \{Good, Bad\}$

$$P(X, Y)$$

	Oranges	Apples	Peaches
Good	0.16	0.28	0.12
Bad	0.12	0.04	0.28

- The sum rule: marginalizing over one r.v. gives the marginal distribution for the other:
  - Discrete case: $P(X = x) = \sum_j P(X, Y_j)$
  - Continuous case: $p_X(x) = \int_{-\infty}^{\infty} p(x, y)dy$

	Oranges	Apples	Peaches	$P(Y)$
Good	0.16	0.28	0.12	0.56
Bad	0.12	0.04	0.28	0.44
$P(X)$	0.28	0.32	0.40	

- Conditioning

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

$$P(Y = \text{bad}|X = \text{apple}) = \frac{P(Y = \text{bad}, X = \text{apple})}{P(X = \text{apple})}$$

	Oranges	Apples	Peaches
Good	0.16	0.28	0.12
Bad	0.12	0.04	0.28

$P(Y = \text{bad}, X = \text{apple})$

- When picking up a random apple, what is the probability it's bad?
  $P(Y = \text{bad}|X = \text{apple}) = \frac{P(Y=\text{bad},X=\text{apple})}{P(X=\text{apple})} = \frac{0.04}{0.32} = 0.125$
- When handed a bad fruit, what is the probability it's a peach?
  ?

	Oranges	Apples	Peaches	$P(Y)$
Good	0.16	0.28	0.12	0.56
Bad	0.12	0.04	0.28	0.44
$P(X)$	0.28	0.32	0.40	

- When picking up a random apple, what is the probability it's bad?
  $P(Y = \text{bad}|X = \text{apple}) = 0.125$
- When handed a bad fruit, what is the probability it's a peach?
  $P(X = \text{peach}|Y = \text{bad})$

	Oranges	Apples	Peaches
Good	4	7	3
Bad	3	1	7

Classification example: Credit scoring

Credit scoring model: Is a potential customer high-risk or not?
We choose two relevant inputs: Income and savings, represented as two random (observable) variables X_1 and X_2

- Input: $\boldsymbol{x} = [x_1, x_2]^T$
- Desired output, high-risk or not: $C \in \{1, 0\}$
- Random variable conditioned on X_1 and X_2

Input: $\boldsymbol{x} = [x_1, x_2]^T$
Desired output: $C \in \{1, 0\}$

We want to know $P(C|X_1, X_2)$ so we can:

$$\text{Choose} = \begin{cases} C = 1 \text{ if } P(C = 1|x_1, x_2) > P(C = 0|x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

with probability of error = $1 - max(P(C = 1|x_1, x_2), P(C = 0|x_1, x_2))$

$$\text{Bayes' rule: } P(C|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C)P(C)}{p(\boldsymbol{x})}$$

- *Prior* probability $P(C)$: how likely it is to observe a class label C, regardless of $\boldsymbol{x}$.
  $p(C) \geq 0$ and $p(C=1) + p(C=0) = 1$
- Class *likelihood* $p(\boldsymbol{x}|C)$: how likely it is that, having observed an example with class label C, the example is at $\boldsymbol{x}$.
  (The distribution of $\boldsymbol{x}$ for each class)

$$\text{Bayes' rule: } P(C|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C)P(C)}{p(\boldsymbol{x})}$$

- *Evidence* $p(\boldsymbol{x})$: probability of observing an example $\boldsymbol{x}$ at all (regardless of its class).
  $p(\boldsymbol{x}) = p(\boldsymbol{x}|C = 0)p(C = 0) + p(\boldsymbol{x}|C = 1)p(C = 1)$
- *Posterior* probability $p(C|\boldsymbol{x})$: how likely it is that, having observed an example $\boldsymbol{x}$, its class label is C.
  $p(C = 1|\boldsymbol{x}) + p(C = 0|\boldsymbol{x}) = 1$
  (This is what we want to know)

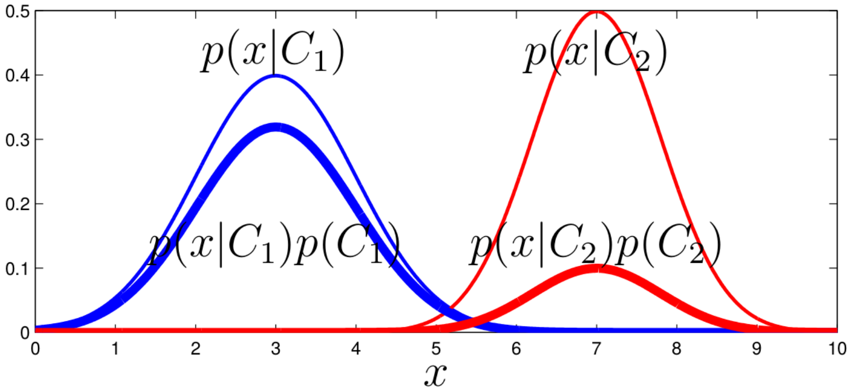
Bayes' rule:

$$P(C|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C)P(C)}{p(\boldsymbol{x})}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

Example, likelihoods and priors

Class likelihoods $p(x|C_k) \sim \mathcal{N}(\mu_k, \sigma_k^2)$



with different prior probabilities $p(C_k)$

http://faculty.ucmerced.edu/mcarreira-perpinan

Example: Disease test results (Ex. 1)

- Rare disease (d) with occurrence 10^{-6}
- Test (t) for disease:
  - 99% chance of finding the disease, if present
  - 10^{-3} chance of wrongly finding the disease, when not present
- A random person is tested, and the result comes out positive:
  What is the probability that this person has the disease?

Example: Disease test results

using 0, 1 encoding for *True, False*

$$
\begin{aligned}
P(d=1|t=1) &= \frac{P(t=1|d=1)P(d=1)}{P(t=1)} \\
&= \frac{P(t=1|d=1)P(d=1)}{P(t=1|d=1)P(d=1)+P(t=1|d=0)P(d=0)} \\
&= \frac{0.99 \cdot 10^{-6}}{0.99 \cdot 10^{-6} + 10^{-3} \cdot (1-10^{-6})} = 0.00098902
\end{aligned}
$$

So still more likely that the person does not have the disease

Generalization to more than two classes:

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{P(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|C_i)P(C_i)}{\sum_{k=1}^{K} p(\boldsymbol{x}|C_k)P(C_k)}$$

where:

$P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

Choose the class with $\text{argmax}_{k=1,\ldots,K}\, p(C_k|\mathbf{x})$

- So far, we've classified according to highest posterior probability
- This minimizes the expected classification error
- But what if some mistakes are more costly than others? E.g. cancer diagnosis, earthquake prediction

- So far, we've classified according to highest posterior probability
- This minimizes the expected classification error
- But what if some mistakes are more costly than others? E.g. cancer diagnosis, earthquake prediction

Choose the class with lowest risk instead

Losses and Risks

Define:

- Actions: α_i
- Loss (cost) of α_i when real class is C_k : λ_{ik}

Example: Loss matrix:

$$\begin{array}{cc} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

Image from: Bishop 2006

- Actions: α_i
- Loss (cost) of α_i when real class is C_k : λ_{ik}
- Expected risk:

$$R(\alpha_i|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\boldsymbol{x})$$

- choose α_i if $R(\alpha_i|\boldsymbol{x}) = \min_k R(\alpha_i|\boldsymbol{x})$

0/1 Loss

A special case is:

$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$$

This gives us:

$$\begin{aligned} R(\alpha_i|\boldsymbol{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k|\boldsymbol{x}) \\ &= \sum_{k \neq i} P(C_k|\boldsymbol{x}) \\ &= 1 - P(C_i|\boldsymbol{x}) \end{aligned}$$

The reject option

- Sometimes it is better to discard uncertain classifications (and leave the final decision to human review or follow-up systems).
- to do that we add 'reject' as an extra action: α_{K+1}

For the 0/1 loss:

$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ \lambda \text{ if } i = K+1,\ 0 < \lambda < 1 \\ 1 \text{ otherwise} \end{cases}$$

$$R(\alpha_{K+1}|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda P(C_k|\boldsymbol{x}) = \lambda$$

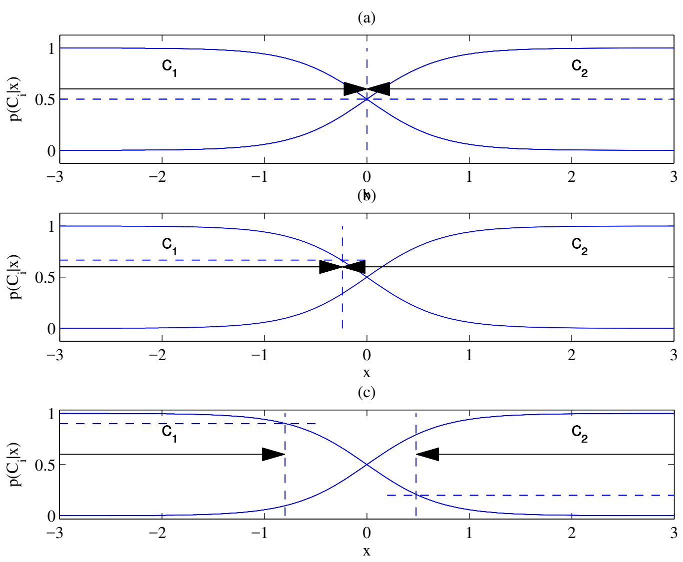
$$R(\alpha_i|\boldsymbol{x}) = \sum_{k \neq i} P(C_k|\boldsymbol{x}) = 1 - P(C_i|\boldsymbol{x})$$

$$\text{choose} \begin{cases} C_i \text{ if } P(C_i|\boldsymbol{x}) > P(C_k|\boldsymbol{x})\ \forall\ k \neq i \text{ and } P(C_i|\boldsymbol{x}) > 1 - \lambda \\ \text{else reject} \end{cases}$$

In extreme cases of λ:

- $\lambda = 0$: always reject (rejecting is less costly than a correct classification).
- $\lambda = 1$: never reject (rejecting is costlier than any misclassification)

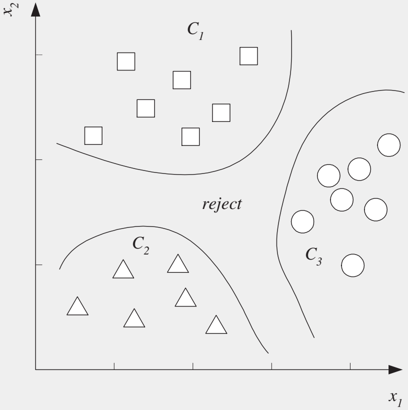
Effect of loss and rejection



- Classification can also be seen as a set of discriminant functions: $g_i(\boldsymbol{x}),\ i = 1, \ldots, K$
- Choose C_i if $g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x})$
- With 0/1 Loss: $g_i(\boldsymbol{x}) = P(C_i|\boldsymbol{x})$
  Ignoring $p(x)$: $g_i(\boldsymbol{x}) = P(\boldsymbol{x}|C_i)P(C_i)$

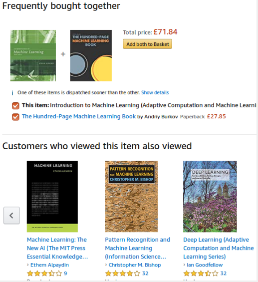
Discriminant Functions

- Discriminant functions divide the feature space into K decision regions $R_1, \ldots, R_K$
- where R_i is the region $\{\boldsymbol{x} | g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x})\}$
- The regions are separated by decision boundaries where there are ties



Association rules

- Association rule: $X \rightarrow Y$ (X: antecedent, Y: consequent)
- E.g.: People who {buy/click/visit} X are also likely to {buy/click/visit} Y
- Used for basket analysis and product recommendations
- $X \rightarrow Y$ implies association, not necessarily causation



Simple but popular measures:

- Support(X, Y)
- Confidence($X \rightarrow Y$)
- Lift($X \rightarrow Y$)

Image source: dlpng.com

$$\text{Support}(X, Y) \equiv P(X, Y) = \frac{\#\text{purchases of } X \text{ and } Y}{\#\text{purchases}}$$

For the rule to be significant, the support should be large. High support means items X and Y are frequently bought together

$$\text{Confidence}(X \to Y) \equiv P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{\#\text{purchases of } X \text{ and } Y}{\#\text{purchases of } X}$$

For the rule to hold with enough confidence, should be $\gg p(Y)$ and close to 1

Basket Analysis - Measures

$$\text{Lift}(X \to Y) = \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)}$$
$$= \frac{\#\text{purchases of } X \text{ and } Y}{(\#\text{purchases of } X) * (\#\text{purchases of } Y)}$$

This is the ratio of the observed joint probability to that expected under independence

- Lift < 1: *X* makes *Y* less likely
- Lift = 1: *X* and *Y* are independent
- Lift > 1: *X* makes *Y* more likely

Basket analysis example (Ex. 3.7)

Given the following data of transactions at a shop, calculate the support and confidence values of milk → bananas, bananas → milk, milk → chocolate, and chocolate → milk.

Transaction	Items in basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

Agrawal et al., 1996

- For (X, Y, Z), a 3-item set, to be frequent (have enough support), (X, Y), (X, Z), and (Y, Z) should be frequent.
- If (X, Y) is not frequent, none of its supersets can be frequent.
- Once we find the frequent k-item sets, we convert them to rules:
  - $X, Y \rightarrow Z, \ldots$
  - $X \rightarrow Y, Z, \ldots$

The Apriori algorithm

- Initially, scan DB once to get frequent 1-itemset
- Repeat
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB to find frequent (k+1) itemsets
  - Set k = k+1
- Terminate when no frequent or candidate set can be generated
- Return all the frequent itemsets derived

Rule Generation

Convert the found itemsets into rules with enough confidence, by splitting the itemset into antecedent and consequent

For $X \rightarrow Y, Z$ to have enough confidence, $X, Y \rightarrow Z$ must have enough confidence

- Start by considering a single consequent and test the confidence for all possible single consequent
- Then consider two consequents for the added rules; etc.

Sources and Resources

- 'Pattern Recogniton and Machine Learning', Bishop, 2006
- probabilitycourse.com
- http://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/index.html
- http://faculty.ucmerced.edu/mcarreira-perpinan/teaching/CSE176/lecturenotes.pdf
- https://slideplayer.com/slide/4778004/ and /6275779/