Course ID 80250993 Dates: 9/16-12/23/2021 @ I-205
Tencent/VooV Meeting 521 4678 6257 (passwd 1205)
Broadcast and Reply: https://meeting.tencent.com/l/p11V4drTYoa4 (passwd 1205)

Part II
Probabilistic Learning Machines

Xuegong Zhang
November 4, 2021

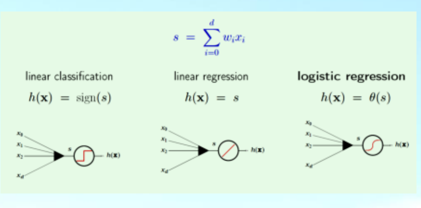
80250993 Machine Learning
@Tsinghua University

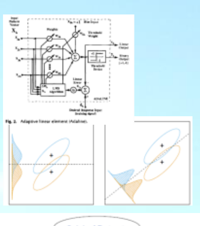
Chapter 10
Bayesian Classifiers

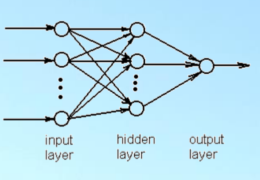
Xuegong Zhang
November 4, 2021

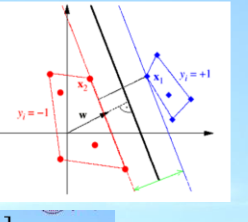
10.1
Probabilistic View of Classification Task

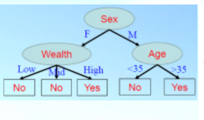
What we have learned

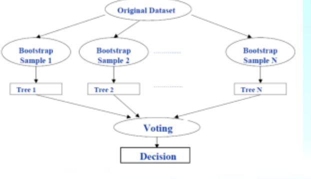


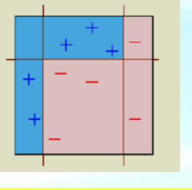


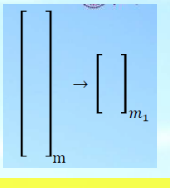












Most methods we have learned so far are of a deterministic manner.

- Output: mostly deterministic (some probabilities)
- Input: fixed training samples
- Computation: deterministic

Probabilistic View: Let's guess the coin

How do we make decision without any observation?

- Two classes: a cent (ω_1) vs. a dime (ω_2)
- How do you decide if we are not allowed to see it?
  - **Guess!**
- What is the base of our guess?
  - Probability: *a prior* probability $P(\omega_1), P(\omega_2)$
    - If $P\ (\omega_1) > P(\omega_2)$, assign $x \in \omega_1$
    - If $P\ (\omega_1) < P(\omega_2)$, assign $x \in \omega_2$

    Why?
  - Error rate: the probability of making a wrong guess

$$Error = \min\{P(\omega_1), P(\omega_2)\}$$

The principle:
Minimize the error!

- The intuition is to make the decision with minimal probability of error.

How do we decide after we weigh it?

- Suppose $x = 2.4\ grams$, is it a cent or a dime?
- What would we do to make an **educated guess**?

United States	
Value	0.01 U.S. Dollars
Mass	2.5 g (0.08 troy oz)
Diameter	19.05 mm (0.75 in)
Thickness	1.52 mm (0.0598 in)
Edge	Plain

United States	
Value	0.10 U.S. dollar
Mass	2.268 g (0.0729 troy oz)
Diameter	17.91 mm (0.705 in)
Thickness	1.35 mm (0.053 in)
Edge	118 reeds

D

- Now what is the base of our guess?
  - a *posterior* probabilities $P(\omega_i|x),\ i = 1,2$
    - If $P\ (\omega_1|x = 2.4) > P(\omega_2|x = 2.4)$, $x \in \omega_1$
    - If $P\ (\omega_1|x = 2.4) < P(\omega_2|x = 2.4)$, $x \in \omega_2$

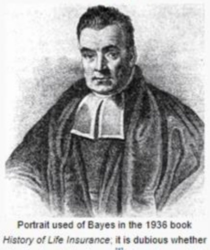
The principle:
Minimize the error!

$$Error = \min\{P(\omega_1|x), P(\omega_2|x)\}$$

How do we know the posterior $P(\omega_i|x)$?

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

Thomas Bayes



Portrait used of Bayes in the 1936 book *History of Life Insurance*; it is dubious whether it actually depicts Bayes.[1] No earlier portrait or claimed portrait survived.

Born	c. 1701 London, England
Died	7 April 1761 (aged 59) Tunbridge Wells, Kent, England
Nationality	English
Signature	T. Bayes.

Bayes' theorem
Bayes' law
Bayes' rule

The conditional density of weight:
-- the model for making the coin

The prior

$$P(\omega_i|x) = \frac{p(x,\omega_i)}{p(x)} = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

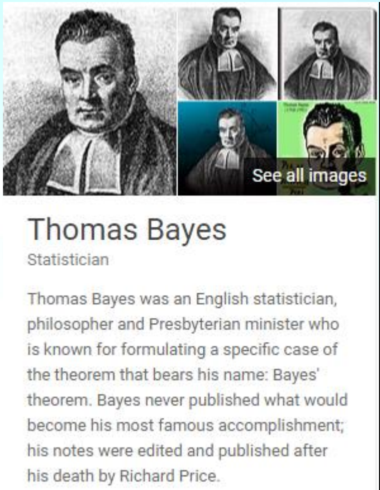
The posterior

The probability density of weight of all coins

Does not affect the comparison of ω_1 and ω_2

W

Everything you ever wanted to know about Bayes' Theorem but were afraid to ask



10.2
Minimal Error Bayes Classifier

Basic terms and symbols

- Samples $\boldsymbol{x} \in \mathrm{R}^d$
- States: $\omega = \omega_1$ for class 1, $\omega = \omega_2$ for class 2
- Prior (*a priori* probability) $P(\omega_1), P(\omega_2)$
- Density (sample distribution density) $p(\boldsymbol{x})$

  (*aka.* pooled probability density)
- Class-conditional density $p(\boldsymbol{x}|\omega_1),\ p(\boldsymbol{x}|\omega_2)$
- Posterior (*a posteriori* probability) $P(\omega_1|\boldsymbol{x}),\ P(\omega_2|\boldsymbol{x})$
- Error rate: probability of error

$$P(e|\boldsymbol{x}) = \begin{cases} P(\omega_2|\boldsymbol{x}), \text{if } \boldsymbol{x} \text{ is assigned to } \omega_1 \\ P(\omega_1|\boldsymbol{x}), \text{if } \boldsymbol{x} \text{ is assigned to } \omega_2 \end{cases}$$

- Average error rate $P(e) = \int P(e|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$
- Probability of correctness $P(c) = 1 - P(e)$

Bayesian decision for minimal error

- Setting of the problem:

  Given the number of classes (states) $\omega_i, i = 1, \cdots, c$, the prior and conditional densities

$$P(\omega_i),\ P(\boldsymbol{x}|\omega_i),\ i = 1, \cdots, c\ ,$$

  find the decision rule that minimize the average error rate, i.e.,

$$\min P(e) = \int P(e|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}.$$

- Solution:
  - Since $P(e|\boldsymbol{x}) \geq 0$ and $p(\boldsymbol{x}) \geq 0$, it is equivalent as to $\min P(e|\boldsymbol{x})$ for all $\boldsymbol{x}$.

    As $P(e|\boldsymbol{x}) = \begin{cases} P(\omega_2|\boldsymbol{x}), & \text{if } \boldsymbol{x} \text{ is assigned to } \omega_1 \\ P(\omega_1|\boldsymbol{x}), & \text{if } \boldsymbol{x} \text{ is assigned to } \omega_2 \end{cases}$, we get the Bayesian Decision Rule (for minimal error)

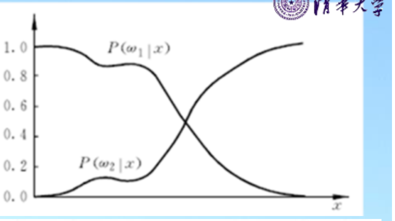
$$\text{If } P(\omega_1|\boldsymbol{x}) \gtrless P(\omega_2|\boldsymbol{x}), \text{assign } \begin{matrix} \boldsymbol{x} \in \omega_1 \\ \boldsymbol{x} \in \omega_2 \end{matrix}$$

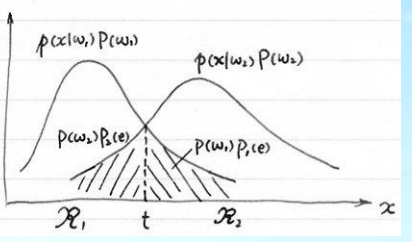
→ Maximum *a posterior* probability (MAP)

Calculation

If $P(\omega_1|\boldsymbol{x}) \gtrless P(\omega_2|\boldsymbol{x})$, assign $\begin{matrix}\boldsymbol{x} \in \omega_1 \\ \boldsymbol{x} \in \omega_2\end{matrix}$

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{2} p(\boldsymbol{x}|\omega_j)P(\omega_j)}, \quad i = 1,2$$





$$P(e) = P(\omega_2)P_2(e) + P(\omega_1)P_1(e)$$

$$= P(\omega_2)\int_{\Re_1} p(\boldsymbol{x}|\omega_2)dx + P(\omega_1)\int_{\Re_2} p(\boldsymbol{x}|\omega_1)dx$$

Equivalent forms of Bayesian Decision Rule

If $P(\omega_1|\boldsymbol{x}) \gtrless P(\omega_2|\boldsymbol{x})$, assign $\begin{matrix}\boldsymbol{x} \in \omega_1 \\ \boldsymbol{x} \in \omega_2\end{matrix}$

① If $P(\omega_i|\boldsymbol{x}) = \max_j P(\omega_j|\boldsymbol{x})$, then $\boldsymbol{x} \in \omega_i$.

② If $p(\boldsymbol{x}|\omega_i)P(\omega_i) = \max_j p(\boldsymbol{x}|\omega_j)P(\omega_j)$, then $\boldsymbol{x} \in \omega_i$.

③ If $l(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \gtrless \frac{P(\omega_2)}{P(\omega_1)}$, then $\boldsymbol{x} \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$.

④ Let $h(\boldsymbol{x}) = -\ln[l(\boldsymbol{x})] = -\ln p(\boldsymbol{x}|\omega_1) + \ln p(\boldsymbol{x}|\omega_2)$,

If $h(\boldsymbol{x}) \lessgtr \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right)$, then $\boldsymbol{x} \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$.

D

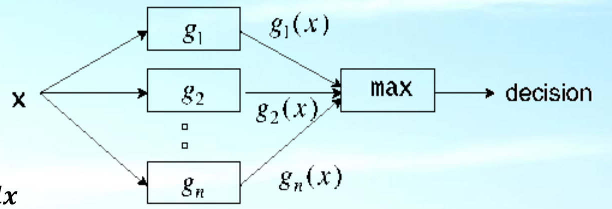
where we call $l(\boldsymbol{x})$ as *likelihood ratio*, $\frac{P(\omega_1)}{P(\omega_2)}$ as threshold of likelihood ratio, and $h(\boldsymbol{x})$ as *log likelihood ratio*.

Bayesian Decision Rule for multi-class cases

If $P(\omega_1|\boldsymbol{x}) \gtrless P(\omega_2|\boldsymbol{x})$, assign $\begin{matrix}\boldsymbol{x}\in\omega_1\\ \boldsymbol{x}\in\omega_2\end{matrix}$

① If $P(\omega_i|\boldsymbol{x}) = \max\limits_{j=1,\cdots,c} P(\omega_j|\boldsymbol{x})$, then $\boldsymbol{x} \in \omega_i$.

② If $p(\boldsymbol{x}|\omega_i)P(\omega_i) = \max\limits_{j=1,\cdots,c} p(\boldsymbol{x}|\omega_j)P(\omega_j)$, then $x \in \omega_i$.



$P(e) = 1 - P(c) = 1 - \sum_{j=1}^{c} P(\omega_j) \int_{\Re_j} p(\boldsymbol{x}|\omega_j) d\boldsymbol{x}$

10-second break

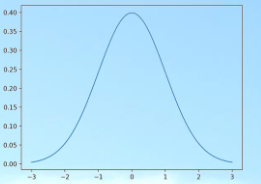
10.3
Examples of Bayes Classifiers

10.3.1
Bayes Classifiers for Normal Distributions

Normal (Gaussian) Distribution

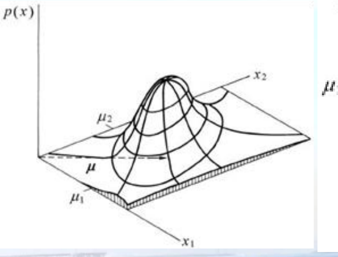
$$N(\mu,\sigma^2):\ \ p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

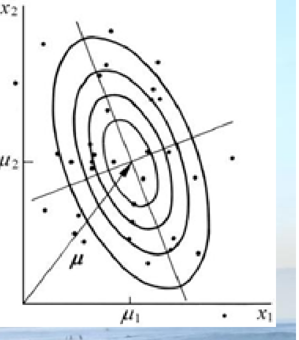
$$\mu=E\{x\}=\int xp(x)dx,\quad \sigma^2=\int(x-\mu)^2p(x)dx=E\{\ (x-\mu)^2\}$$



$$N(\boldsymbol{\mu},\boldsymbol{\Sigma}):\ \ p(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\quad \boldsymbol{x}\in R^d$$

$$\boldsymbol{\mu}=E[\boldsymbol{x}],\ \Sigma=E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}]$$





Bayesian Decision with Gaussian Distribution: the general case

$$N(\boldsymbol{\mu},\boldsymbol{\Sigma}):\ \ p(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\quad \boldsymbol{x}\in R^d$$

Conditional Density $\quad p(\boldsymbol{x}|\omega_i)=\frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]$

Discriminant $\quad g_i(\boldsymbol{x})=\ln[p(\boldsymbol{x}|\omega_i)P(\omega_i)]=\ln p\,(\boldsymbol{x}|\omega_i)+\ln P\,(\omega_i)$

Decision boundary $\quad g_i(\boldsymbol{x})=g_j(\boldsymbol{x})$

$$-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]-\frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}+\ln\frac{P(\omega_i)}{P(\omega_j)}=0$$

--- Quadratic discriminant

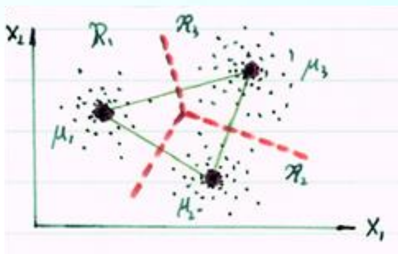
Some special cases:

- $\Sigma_i = \sigma^2 I,\ i = 1, \cdots, c,$ and all priors $P(\omega_i),\ i = 1, \cdots, c$ are equal

$$g_i(\boldsymbol{x}) = \ln[p(\boldsymbol{x}|\omega_i)P(\omega_i)] = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$$g_i(\boldsymbol{x}) = -\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_i) = -\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2 = -\frac{1}{2\sigma^2}\sum_{j=1}^{d}(x_j-\mu_{ij})^2$$

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x} + b_i,$$
$$\boldsymbol{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i,$$
$$b_i = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i$$



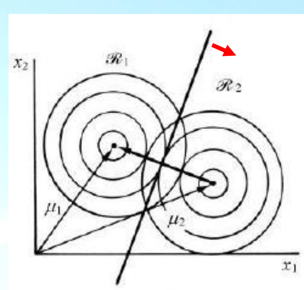
- Minimal distance classifier
- Template-matching

Some special cases:

- $\Sigma_i = \sigma^2 I,\ i = 1, \cdots, c,$ but priors $P(\omega_i),\ i = 1, \cdots, c$ are NOT equal

$$g_i(\boldsymbol{x}) = \ln[p(\boldsymbol{x}|\omega_i)P(\omega_i)] = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x} + b_i,$$
$$\boldsymbol{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i,$$
$$b_i = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i + \ln P(\omega_i)$$



- The decision boundary will move toward the class with smaller prior.

D

Some special cases:

- $\Sigma_i = \Sigma,\ i = 1, \cdots, c$

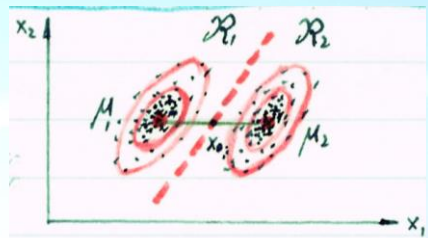
$$g_i(\boldsymbol{x}) = \ln[\,p(\boldsymbol{x}|\omega_i)P(\omega_i)] = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| - \boxed{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)} + \ln P(\omega_i)$$

Mahalanobis Distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$

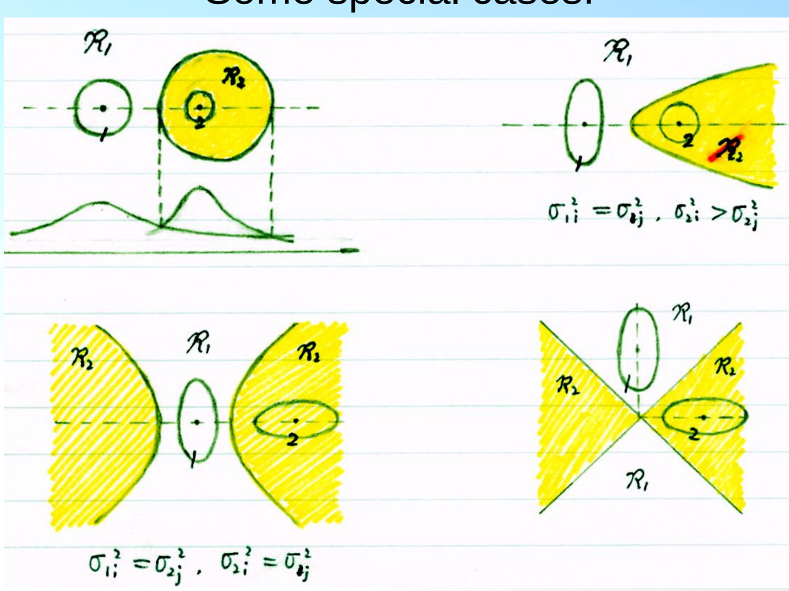
$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x} + b_i,$$

$$\boldsymbol{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i,$$

$$b_i = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

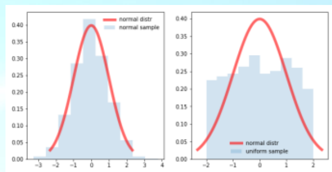


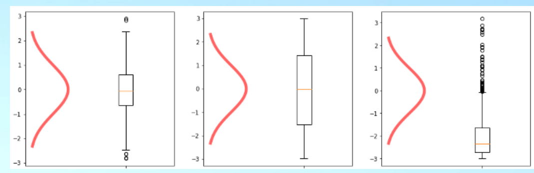
Some special cases:



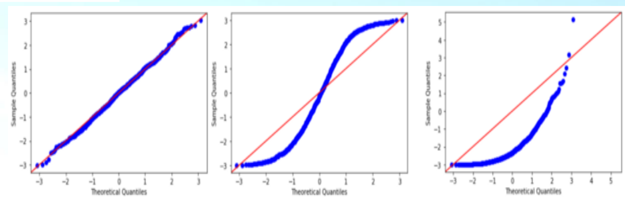
How do I know whether my data are Gaussian?

- ✓ Reasoning
  - – Very likely to be Gaussian if the quantity is affected by many factors with no one dominating
- ✓ Visual checking
  - – Histograms
  - – Boxplots
  - – QQ-plots





- ✓ Statistical tests
  - – χ^2 goodness-of-fit test
  - – Kolmogorov Smirnov test (KS test)
  - – Lilliefors test
  - – Shapiro Wilk test
  - – …



https://towardsdatascience.com/6-ways-to-test-for-a-normal-distribution-which-one-to-use-9dcf47d8fa93

单选题 1分 设置

4-min break

- A Yes, I’m back to my computer.
- B Sorry, not yet.

提交

10.3.2
Bayes Classifiers for Discrete Sequence Data

An example in genomic sequence analysis

```
13141 atttacaaca acaacaacca aaaataaaaa cacatacaca caagtgatag cagcagacaa
13201 cacaaaacat aacaaaaaca gaagcttccc aaaattccga aattcaaaat cctctttaat
13261 ttctcatcat atagtagtcg agttacgttg gcatggtttc taatagttat tatgcggagt
13321 tttaaatctt catcagttgt attctttcga attttgcgaa ctcagaatcg actgcaaata
13381 aacatgttta attagatcaa atggttagag attccgagct tgccttgtaa catgacatga
13441 caacaatgat ttggctataa attacttgta agttactagt gagaaacata gcaaaccaat
13501 tttcaagtat accacagtgt tagttttgac taacattagg tctaattcgg tgaagtcaat
13561 actaatccaa agtttcatct atcaatgcga atagatttaa aattattgag gagcacaaac
13621 agattcaaaa tgttaaatca ttcaaaaaca tttagtaata gagcaaatta ttatgcattt
13681 accatatttc atttttgtg ctacagtata tccacaaatt aaaaagtgat gtgattaata
13741 catacettca ctttaggata agggaagcta gttttagcac cacctctgct tccgaaactg
13801 aactttgatt tcttgtcatt agattttaag atctgaaaag agcaggtcaa atcctcagac
13861 acactcatca tcgcacctgc attgtcaaat tccccacagt aattcggcgc agagaatatc
13921 gttacgagct gcttattcgc aaagaactcg aatccatctt caacaacctg ttacaaaatc
13981 gttgcttgtg tcaaaaggaa acacgttaaa aatgtttctt tgggcacaag gaacaataaa
14041 caaacctggt gagccctaca aatgaggtca agatcgagtc ttttaagaaa tccagaaact
14101 atatctgatc caaaagtgta agaaactccg cgatcgttag gcccccaacc tctaacatct
14161 ttatcaggat cagaccacaa gagatcacag agtaaaccac gatcaggaat atccgttgga
14221 cgacgaatat ccctaatctg cctcaagctt agcagctccg gggagagccc accatgcata
14281 caaaaaatcc gctcatcgat gagtgcagcg acggggagac agttgaagca atcagtgaag
14341 attcgccaaa tcttgacact gaatctacgt ttacactcgt catagaagcc gtaaatacga
14401 ttgattgatg cactttcatg gtttcctctg agaaggaaga agttttcagg gaacttaatc
14461 ttgtaagcaa gtaaaagaca aatcgtttcg aggctttgct tgccgcgatc gacataatct
14521 ccaagaaaca agtagtttga attaggaggg tatccgccat gttcgaatag tctcaagaga
14581 tccggatatt gtccatgaat gtcccctgtt tattatttgt ttaaccaaaa gattacatag
14641 gagcacaatc tcagaaaaga gtcagagaga aacagagaga aacgtaccac atattttaac
14701 aggagcttca agttccaaga gatttggttg tctcaagaag atatctctag agacgaaaca
14761 gagctgtttg atctctgttt cagacaactg aacaatcttt cctggttttt ctctagcttc
14821 aagcaaccta ttgatcaccg agtttaaagt accaggatcc atttataaaa cctcaagaaa
14881 cagagcaaca atgagacaaa acaacaacaa cgcctagaag ctttataata tacaaaagct
14941 attgacaata tggaataaat ttaaaaagta tggagctttt atgttctttt atgcttgctc
15001 tttgagaaaa agaatggtga ttgtcaaatt caagaagggc tttataaagc cgcgtttaag
15061 aatatattct tcatacettt ttagcaaaag tcacatgtgt tcatacctac ctacattgtt
15121 ctaacttcaa ctaaatcata taataatcag attctagcta atttagtaac acattgtttg
```

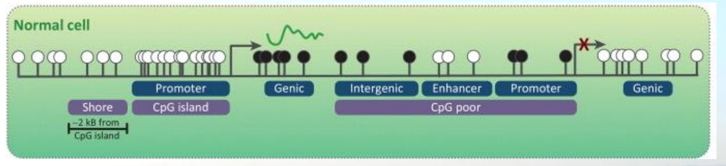
Example: “CpG islands” in the human genome

- In the human genome, the dinucleotide CG occurs rarer (~1/5) than would be expected from the independent probabilities of C and G.
  - Wherever CG occurs, the C is typically chemically modified by methylation, which results in a high chance of this methyl-C mutating into a T.

→ Less “CG” dinucleotide (called CpG) observed

- The methylation process is suppressed in short stretches around the promoters ("start" regions) of many genes.
  - In these regions, we see many more CpG dinucleotides than elsewhere.
  - Such regions are called CpG islands (typically a few hundred to a few thousand bases long).

A CpG site, i.e., the " 5'—C—phosphate—G—3' " sequence of nucleotides, is indicated on one DNA strand (in yellow). On the reverse DNA strand (in blue), the complementary 5'—CpG—3' site is shown. A C-G base-paring between the two DNA strands is also indicated (right)



Exon 1 CpG Island: 12634..12767

```
 11941 ttataagatc cccctccctc taaatcctgt ccttctatca cttcatcctt CGctctcctt
 12001 taaaatgaga cagttgtcag caggaatcct gCGcaagaac acaccaccct gtttcataga
 12061 agatatctca ggtaatgtgc aaacaCGggt ttttaaaCGg agCGcatttt tctcatttgt
 12121 taatatcacc acctaaatca tctcttgcct aaaacaagga gtagaaagtg aatgaaggaa
 12181 ggaacaggtg atggtcagtg tcctttctac gcctcaaaat ttaagagttt atgtgaaaat
 12241 tcataaatat taatctcaat ccaggttaag caaaattttt tgctctcctc tttagaaatt
 12301 tctggttgcc aaagttccag aaattgcttc ctcattcctg agcctttcat tttctCGatt
 12361 tctccattat gtaaCGggga gctggagctt tgggcCGaat ttccaattaa agatgatttt
 12421 tacagtcaat gagccaCGtc agggagCGat ggcaccCGca ggCGgtatca actgatgcaa
 12481 gtgttcaagc gaatctcaac tCGtttttttc CGgtgactca ttccCGgccc tgcttggcag
 12541 CGctgcaccc tttaacttaa acctCGgcCG gcCGccCGcc gggggcacag agtgtgCGcc
*12601 gggcCGCGCG gcaattggtc ccCGCGcCGa cctcCGccCG CGagCGcCGc CGcttccctt
*12661 cccCGcccCG CGtccctccc cctCGgcccc gCGCGtCGcc tgtcctcCGa gccagtCGct
*12721 gacagcCGCG gCGcCGCGag cttctcctct cctcaCGacc gaggcaggta aaCGccCGgg
 12781 gtgggaggaa CGCGggCGgg ggcaggggag cCGCGggggc CGagtgagga cccCGggcct
 12841 CGggtcccag gCGcaagggt gccCGgcCGg gCGgggtCGg gaccccagtg aggaggggcc
 12901 gggggctgcc cCGCGggCGc gtgaCGgtct CGggcctgcc CGgctgCGct ggtctcCGct
 12961 CGggtgaggc ggcttggctt CGcttttcag gttaggaaag ctccctttac tgCGCGttgg
 13021 ggggctgggg gagctggCGg agccaCGtta gggaggtCGg tggCGcCGgg gtgtctcagc
 13081 gccccctgca cccCGCGCGg gtcCGgccca gCGggCGatc gctggCGccc agggaactcc
 13141 gggagggcCG ccagCGggct cCGcaggCGc ggggCGggga ggggCGcctg ggggcCGCGg
 13201 ggctCGCGct cccCGccCGt tggcCGcccc tCGgaggcCG agatCGgggc ccagaaCGcc
 13261 ccttggcaaa gcctggCGct tcCGCGatgc ccagagggtg cttgggggga tggagagagg
 13321 ggCGccCGcc ggggtagttc CGggagcctc ggtgcctccc gcCGcagctg cagCGttcct
 13381 ccCGggaggc ggcccagccc ttcatcctCG cCGcctgagc ttctcCGagg ggggctgcag
 13441 ccttgCGgcc gttgccacCG cctggagaag CGgcccaCGc ggactgaCGg gCGggggCGg
 13501 ggcctCGggc ctCGgCGggg gCGgggtcCG gggaggcccc accctctgtt ctccaggggc
 13561 ggggagagag gagctgcagg tctgCGgcct ggccccaggt gCGatggCGg accccagctt
 13621 ggccagtcac attcctccca gtccccctgg agggagaaCG ctggccatgg ggggctccaa
 13681 ggaacaacca gcctCGgatg aCGacccttg ggtcacCGgt ctccccacct gtgCGgcagg
 13741 CGccttcaCG tttcattatt aaacaatggg gagaaatcca tgtttactgt ccttttttagg
 13801 aattttttgc tcttctcttt gaggtggctg taggaaatag attttttttt taacctCGca
 13861 attccaccac ggtcacatcc atcctCGcca tCGcagagcc acagctctcc gtttttgttt
 13921 cctagcctcc agattctcac acaacacagt gcagtttcac tgctgtaatg atgaggatct
 13981 tcatggcCGc gttattttct tgttctgaga gcatcaCGgt ttaattagca gttccccata
 14041 tgatttgaag tgtttccCGt ttccttaggg aaaactcctg gtagaatagg attaaggatt
 14101 tttacaaata taattatcaa aaacatagga acagggaatt ggataaatat gttaaacttc
 14161 tggaaaaatc aacaaCGctc ttagatttgt agaagaaagg aaaaaatcac cagtggaaag
 14221 gagcaatttt acttacacaa acacagagaa ggtcttacag tgaaaaaaag ctaaccagta
```

http://www.mad-cow.org/exon2.html

```
GGAGGGGCTG GGGGTCCCCA GGACTGCATG CGCCAGGCAG ACGGAGGGGA GGCCGAGCCA
GGAGCCGCAG GCTTCTCAGT GCCAGACCAA CGTCGCCCTC CGCTGGCCGT GTGGGGAATC
GCGCCTAGGC TGGGGCTCCC CTGGCCCCAG GTTAGGGGGA CTCCGATGTG TCCTGGGGCG
CTTCCTGCCC ATCCAGCCTC CGCCAGGCAG CCTGGCCCTG AATGGCTTTG AGGGCAGCCT
CCGCAGCACA GGACCTGCCT CCGCCCTGTG GGGCAGATTT GGGCGTCTGA CTCTGGACAC
TTACCCAGAG GCTGAATTCA ACTCCCCAGC GAAGAGCACA CTCAGACTCC CTCCCCGCTG
TGCCCAGCCT GAGGGGACCC TGGCTTGGAC TTGGGAGGAG AAGGCCAGAG GCCAGGGCCT
GGCTCTCAGG CACTAGGTGC AGGAGAAAAT GGAGCAGAGT TCCTGCAGGA ACTGCCTCAG
ACTGCCCCCG AGGGGACCTG GCCTGGAACC CTCCCTGTGC TACCAGGCTG CTGTTGGGGC
CAGGAGGGCA GGGAGAGCAG CCACCTGTAC GGCATCAAGT AACACTATCA GTCAGCACAC
CCAACACCCA CACGACAGCC ACAGCCCTCT GCCCTGAGGT GGCTGCAACT GAGCAGAGTC
TCCTGTCTTG TTCACCCTGA GCGCGGAGCT CCGCACAGGA CAGGATCAGA GCGCAAACCA
GGAACAAATG GATGAACCAA GGAAGGGGCC AGGGCAGTAG GTCCCCACCC ACACCCTGAA
CCCTAAGGGT GGGACTCAGA GGCTGGAGAC GGGCTGGCCC CAGGATCCTC TAAGTGACAG
GAACCCTTGG CCACTGTTGA TTGCCCCGAG GCGGGGACAG GGATGGTGGG GCTGGGGCTG
CCTGGCCCTT TAAGAGGGCA GTTCTGCCCA CCACCCTGAA TGCCCCTCCC CTAGCTACTG
GAGAAGAGAT CCCGTTTCTT GGCAACAGGA AGCTCTTGGT TTACACAGTG TCACCCAAGC
GACTGGGAGC CGCGTCCTGC TCTGGGACTG AGCCGTTGGA GCTGCCCCGC TGACCGCACT
AGGCTGACCG CAACCGGCTG ACCACACACA GTCCTCACTC CCCTGGCCCT GGTGGGCGGC
AGGCACCATG GCGCAGACAG TGCCGCCCTG CGAGCTGCCC TGCAAAGAGT ACGACGTGGC
```

http://bio.kuas.edu.tw/methyl-typing/userManual.jsp

Deterministic Definition of CpG islands

(Gardiner-Garden & Frommer, 1987)

- Size >200bp
- %(G+C) > 50%

$$\%(\mathrm{G}+\mathrm{C}) = \frac{N(\mathrm{C}) + N(\mathrm{G})}{N}$$

- CpG ratio (Obs/Exp) > 0.60

$$\text{CpG ratio} = \frac{\frac{N(\mathrm{CpG})}{N}}{\frac{N(\mathrm{C})}{N} \times \frac{N(\mathrm{G})}{N}}$$

Markov Chain

- What kind of **probabilistic model** may we use for CpG island regions?
  - Dinucleotides
  - The probability of a symbol depends on the previous symbol
  - ➢ The simplest model is the Markov Chain
- Time series of symbols $x_1, x_2, \cdots, x_n$
- The value at a time point is one of an alphabet (state set)

  {A, C, G, T} for DNA sequences
- 1st-order Markov Chain:

  The time i value depends only on the time $i-1$ value.

$$a_{st} = P(x_i = t | x_{i-1} = s)$$

- Definition:

  $X_1, X_2, \cdots, X_n, \cdots$ is called a (1st order) Markov chain if
  $P(X_n = x_n | X_1 = x_1, \cdots, X_{n-1} = x_{n-1}) \;=\; P(X_n = x_n | X_{n-1} = x_{n-1})$

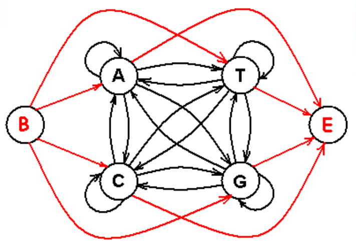
- Transition Probability:

$$a_{st} = P(x_i = t | x_{i-1} = s)$$

The model

- Probability of an instance x of a Markov chain model:

$$P(x) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$



Begin and end states can be added to a Markov chain for modelling both ends of a sequence.

Durbin et al, Biological sequence analysis, pp.48

Recap: Bayesian Decision

$$\text{If } P(\omega_1|\mathbf{x}) \underset{<}{>} P(\omega_2|\mathbf{x}), \text{assign } \begin{matrix}\mathbf{x} \in \omega_1 \\ \mathbf{x} \in \omega_2\end{matrix}$$

$$\text{If } l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{<}{>} \frac{P(\omega_2)}{P(\omega_1)}, \text{ then } x \in \begin{cases}\omega_1 \\ \omega_2\end{cases}.$$

$$\text{If } \; h(x) = -\ln[l(x)] \underset{>}{<} \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right), \text{ then } x \in \begin{cases}\omega_1 \\ \omega_2\end{cases}.$$

Denote CpG island as class + and non-CpG island as class –.
We can use the log likelihood ratio

$$S(x) = \log \frac{P(x|+)}{P(x|-)}$$

to make decision if we can build Markov models of the two classes to calculate $P(x|+)$ and $P(x|-)$.

Using Markov chains for discrimination

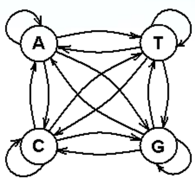
Building models for CpG island (+) and non-CpG island(–) region.

$$a_{st}^{+} = \frac{c_{st}^{+}}{\sum_{t'} c_{st'}^{+}} \qquad a_{st}^{-} = \frac{c_{st}^{-}}{\sum_{t'} c_{st'}^{-}}$$

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

To use the models for discrimination on x, calculate the log-odds ratio:



$$S(x) = \log\frac{P(x|+)}{P(x|-)} = \sum_{i=1}^{L} \log\frac{a_{x_{i-1}x_i}^{+}}{a_{x_{i-1}x_i}^{-}} = \sum_{i=1}^{L} \boldsymbol{\beta}_{x_{i-1}x_i}$$

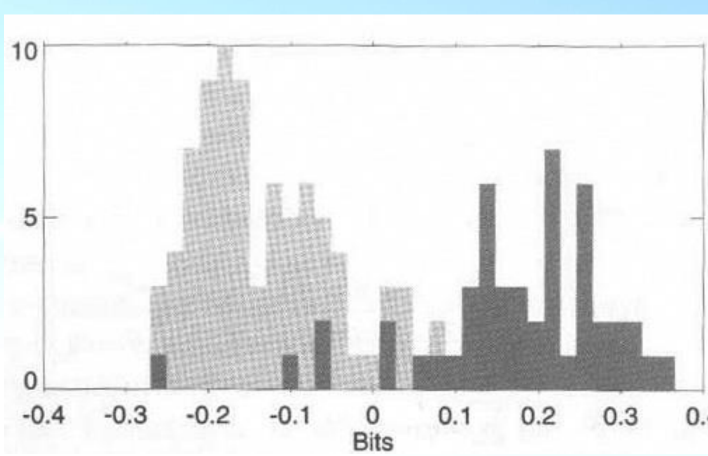
$\boldsymbol{\beta}_{st}$: log likelihood ratios from state s to t

If $l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \gtrless \lambda$, then $x \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$

If $S(x) \gtrless \lambda$, then $x \in \begin{cases}\text{CpG island} \\ \text{non} - \text{CpG island}\end{cases}$

beta	A	C	G	T
A	-.740	0.419	0.580	-.803
C	-.913	0.302	1.812	-.685
G	-.624	0.461	0.331	-.730
T	-1.169	0.573	0.393	-.679

Example: histogram of the length-normalized scores of training sequences



Dark grey: CpG islands
light grey: non-CpG regions

$$S(x) = \log\frac{P(x|+)}{P(x|-)} = \sum_{i=1}^{L} \log\frac{a_{x_{i-1}x_i}^{+}}{a_{x_{i-1}x_i}^{-}} = \sum_{i=1}^{L} \boldsymbol{\beta}_{x_{i-1}x_i}$$

Durbin et al, Biological sequence analysis, pp.52

3-second break

10.4
Naïve Bayes Classifier

For high-dimensional features $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^{\mathrm{T}}$, dealing with $p(\boldsymbol{x}|\omega_i)$ is challenging.

D

H

➢ Naïve Bayes Classifier 朴素贝叶斯分类器

--- Naïve assumption of conditional independence of elements of $\boldsymbol{x}$

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} \propto P(\omega_i)p(x_1, x_2, \cdots, x_d|\omega_i)$$

$$= P(\omega_i)\prod_{k=1}^{d} p(x_k|\omega_i)$$

10-second break

10.5 Minimal Risk Bayes Classifiers

Recall: errors are not equal

- Positive: something (e.g., a disease) is there
- Negative: the thing is not there
- Diagnostic test: using a test to judge positive or negative
- Test performances:
  - Sensitivity
    $P(T^+|D^+) = TP / (TP + FN)$
  - Specificity
    $P(T^-|D^-) = TN / (TN + FP)$
  - Prevalence: pre-test probability (probability in the population)
    $D^+/(D^+ + D^-)$
  - Predictive value of a test
    - Discovery rate: $P(D^+|T^+) = TP / (TP + FP)$
    - False discovery rate: $P(D^-|T^+) = FP / (TP + FP)$
  - Accuracy: $(TP + TN)/(D^+ + D^-)$

	Disease present D^+	Disease absent D^-
Test positive T^+	TP: True Positives	FP: False Positives
Test negative T^-	FN: False Negatives	TN: True Negatives

➢ We should care more about the cost of errors!

The Decision Problem

- Samples: random vectors $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T$
- State space of c possible states: $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$
- Decision space of k possible decisions $\mathcal{A} = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ for sample $\boldsymbol{x}$
- The loss with decision α_i on $\boldsymbol{x}$ of true state ω_j:

$$\lambda(\alpha_i, \omega_j), i = 1, \cdots, k, j = 1, \cdots, c.$$

All possible losses are usually defined as a Decision Table:

α \ λ \ ω	ω_1	$\cdots$	ω_c
α_1	$\lambda(\alpha_1, \omega_1)$	$\cdots$	$\lambda(\alpha_1, \omega_c)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
α_k	$\lambda(\alpha_k, \omega_1)$	$\cdots$	$\lambda(\alpha_k, \omega_c)$

How should we make the decision for sample $\boldsymbol{x}$?

- Expected loss if we choose α_i:

$$R(\alpha_i|\boldsymbol{x}) = E[\lambda(\alpha_i, \omega_j)|\boldsymbol{x}] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|\boldsymbol{x}), \quad i = 1, \cdots, k$$

- The expectation of loss for all samples under decision rule $\alpha(\boldsymbol{x})$:

$$R(\alpha) = \int R(\alpha(\boldsymbol{x})|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

--- Expected risk, or average risk

$R(\alpha)$ is a functional: the function of decision function $\alpha(\boldsymbol{x})$

- Minimal Risk Decision: $\min_{\alpha(\cdot)} R(\alpha)$
- Decision Rule: If $R(\alpha_i|\boldsymbol{x}) = \min_{j=1,\cdots,k} R(\alpha_j|\boldsymbol{x})$, then $\alpha = \alpha_i$

If $R(\alpha_i|\boldsymbol{x}) = \min_{j=1,\cdots,k} R(\alpha_j|\boldsymbol{x})$, then $\alpha = \alpha_i$

How can we get $R(\alpha_i|\boldsymbol{x})$?

$$R(\alpha_i|\boldsymbol{x}) = E[\lambda(\alpha_i, \omega_j)|\boldsymbol{x}] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|\boldsymbol{x}), \quad i = 1, \cdots, k$$

- Posterior

$$P(\omega_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_j)P(\omega_j)}{\sum_{i=1}^{c} p(\boldsymbol{x}|\omega_i)P(\omega_i)}, \quad j = 1, \cdots, c$$

- Risk

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|\boldsymbol{x}), \quad i = 1, \cdots, k$$

- Decision

$$\alpha = \underset{i=1,\cdots,k}{\operatorname{argmin}} R(\alpha_i|\boldsymbol{x})$$

Xuegong Zhang 45

For two classes

If $\lambda_{11}P(\omega_1|\boldsymbol{x}) + \lambda_{12}P(\omega_2|\boldsymbol{x}) \lessgtr \lambda_{21}P(\omega_1|\boldsymbol{x}) + \lambda_{22}P(\omega_2|\boldsymbol{x})$, then $\boldsymbol{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

Loss	ω_1	ω_2
α_1	λ_{11}	λ_{12}
α_2	λ_{21}	λ_{22}

Equivalent forms:

If $$(\lambda_{11} - \lambda_{21})P(\omega_1|\boldsymbol{x}) \lessgtr (\lambda_{22} - \lambda_{12})P(\omega_2|\boldsymbol{x})$$

or $$\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\omega_1)P(\omega_1)}{p(\boldsymbol{x}|\omega_2)P(\omega_2)} \gtrless \frac{\lambda_{22}-\lambda_{12}}{\lambda_{11}-\lambda_{21}} = \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}$$

or

$$l(x) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \gtrless \frac{P(\omega_2)}{P(\omega_1)} \cdot \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$$

Threshold of the likelihood ratio

then $\boldsymbol{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$

When $\lambda_{11} = \lambda_{22} = 0,\ \lambda_{12} = \lambda_{21} = 1$,

minimal risk decision becomes minimal error decision.

Xuegong Zhang 46

Example:

1. Two classes of cells: normal $\mathrm{P}(\omega_1) = 0.9$, disease $\mathrm{P}(\omega_2) = 0.1$. For a cell x, from disease model (knowledge) we got $p(x|\omega_1) = 0.2$ and $p(x|\omega_2) = 0.4$, decide whether x is disease or normal by the Bayesian minimal error decision principle.

- Solution:

$$P(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{\sum_{i=1}^{2} p(x|\omega_i)P(\omega_i)} = \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.4 \times 0.1} = 0.818$$

$$P(\omega_2|x) = 1 - P(\omega_1|x) = 0.182$$

$$0.818 > 0.182, \quad \therefore\ x \in \omega_1$$

2. The same data, but use the given decision table, decide whether $\boldsymbol{x}$ is disease or normal by minimal risk decision.

Loss	ω_1	ω_2
α_1	$\lambda_{11} = 0$	$\lambda_{12} = 6$
α_2	$\lambda_{21} = 1$	$\lambda_{22} = 0$

- Solution:

$$R(\alpha_1|x) = \sum_{j=1}^{2} \lambda_{1j} P(\omega_q|x) = \lambda_{12} P(\omega_2|x) = 6 \times 0.182 = 1.092$$

$$R(\alpha_2|x) = \lambda_{21} P(\omega_1|x) = 1 \times 0.818 = 0.818$$

$$1.092 > 0.818, \quad \therefore\ x \in \omega_2$$

11.6
Minimizing one type of error conditional on the other

Neyman-Pearson Criterion:

to minimize the error on one class under a constrain for the error on the other class

$$\min P_1(e), \qquad \text{s.t. } P_2(e) = \epsilon$$

Lagrangian

$$L = P_1(e) + \lambda(P_2(e) - \varepsilon) = \cdots = (1 - \lambda\varepsilon) + \int_{R_1} [\lambda p(x|\omega_2) - p(x|\omega_1)]dx$$

R_1 is the decision region of class ω_1

Solution:

At the boundary t^* of R_1: $\frac{\partial L}{\partial t^*} = 0$ ➔ $\lambda^* = \left.\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)}\right|_{\boldsymbol{x} \in t^*}$

Decision rule:

$$\text{If } \lambda p(x|\omega_2) \begin{matrix} < \\ > \end{matrix} p(x|\omega_1), \quad \text{then } x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$$\text{If } \quad l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \begin{matrix} > \\ < \end{matrix} \lambda, \quad \text{then } x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$$\text{If } \quad l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \begin{matrix} > \\ < \end{matrix} \lambda, \quad \text{then } x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

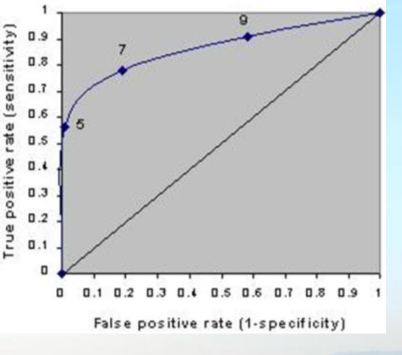
How to solve for the threshold?

Decision region of ω_2 in the λ space $[0, \lambda)$

$$\therefore \quad P_2(e) = 1 - \int_0^{\lambda} p(l|\omega_2)dl$$

- Numeric solution for $P_2(e) = 1 - \int_0^{\lambda} p(l|\omega_2)dl = \epsilon$

- Or, find the proper threshold λ on the ROC curve!



10-second break

10.7
The importance of the prior and the Minimax criterion

Dilemma at the movies

This person dropped their ticket in the hallway.

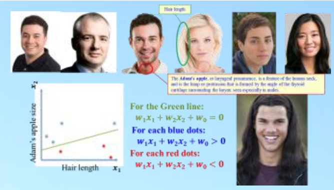
Do you call out

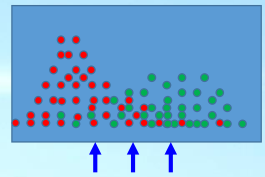
"Excuse me, ma'am!"

or

"Excuse me, sir!"

You have to make a guess.

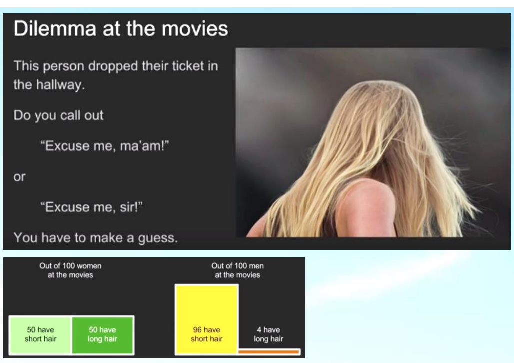


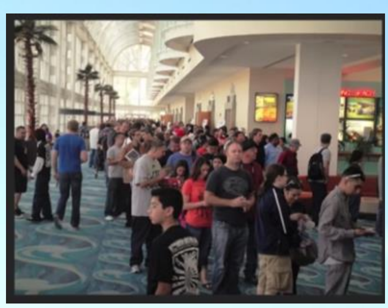


$$P(\textit{sex}|\textit{hair}) = \frac{P(\textit{hair}|\textit{sex})P(\textit{sex})}{\sum_{\textit{sex}} P(\textit{sex})P(\textit{data}|\textit{sex})}$$

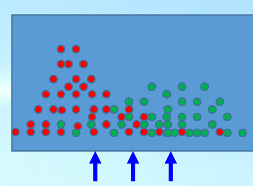
- If $P(M) = P(W)$,
  - $P(M|\text{longhair}) = 0.04A$
  - $P(W|\text{longhair}) = 0.5A$
  - ➢ Woman

Brandon Bohrer: a walk through a couple of Bayesian inference examples.

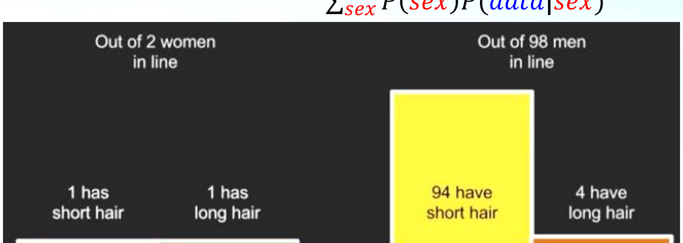




- What if this happens when you are standing in the line for the **Men's Restroom**?



$$P(\textit{sex}|\textit{hair}) = \frac{P(\textit{hair}|\textit{sex})P(\textit{sex})}{\sum_{\textit{sex}} P(\textit{sex})P(\textit{data}|\textit{sex})}$$



- If $P(M) = 48P(W)$ in the line
  - $P(M|\text{longhair}) = 1.92B$
  - $P(W|\text{longhair}) = 0.5B$
  - ➢ Man!

Brandon Bohrer: a walk through a couple of Bayesian inference examples.

The minimax criterion

What shall I do if I don't know the prior?

- Let's look into the risk

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|\boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) \frac{p(\boldsymbol{x}|\omega_j) P(\omega_j)}{\sum_{i=1}^{c} p(\boldsymbol{x}|\omega_i) P(\omega_i)}, \quad i = 1, \cdots, k$$

$$R(\alpha) = \int R(\alpha(\boldsymbol{x})|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

Loss	ω_1	ω_2
α_1	λ_{11}	λ_{12}
α_2	λ_{21}	λ_{22}

Two-class scenario:

$$R = \int_{R_1} [\lambda_{11} P(\omega_1) p(x|\omega_1) + \lambda_{12} P(\omega_2) p(x|\omega_2)\,] dx + \int_{R_2} [\lambda_{21} P(\omega_1) p(x|\omega_1) + \lambda_{22} P(\omega_2) p(x|\omega_2)\,] dx$$

- When the priors $P(\omega_1)$ and $P(\omega_2)$ are unknown

D

  - **Minimax criterion:**
    - ➢Consider the worst case: minimize the overall risk for the worst prior

      ➔ **To minimize the maximum possible overall risk.**
  - Take $P(\omega_1)$ as a variable, and optimize the risk $R\left(P(\omega_1)\right)$ as its function.

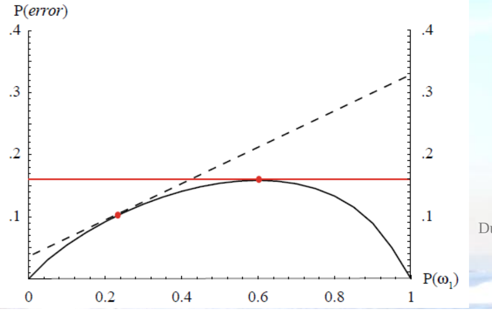
$$R = \int_{R_1} [\lambda_{11}P(\omega_1)p(x|\omega_1) + \lambda_{12}P(\omega_2)p(x|\omega_2)]dx + \int_{R_2} [\lambda_{21}P(\omega_1)p(x|\omega_1) + \lambda_{22}P(\omega_2)p(x|\omega_2)]dx$$

Since $P(\omega_2) = 1 - P(\omega_1)$ and $\int_{R_1} p(x|\omega_1)dx = 1 - \int_{R_2} p(x|\omega_1)dx$, we have

$$R(P(\omega_1)) = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx$$
$$+P(\omega_1)\underbrace{\left\{(\lambda_{11} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)dx - (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx\right\}}_{=0 \text{ for minimax solution}}$$

$$R_{\text{minimax}} = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx = \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)dx$$

- It can be hard to solve for an analytical solution for complicate data, but the idea is useful.



Duda et al, *Pattern Classificaiton*, p.29

Discussion

- Question
  - What is the key difference of Bayesian Decision and the ML methods we learned before (e.g., NN, SVM, …)?
- Bayesian Decision:
  - Model-based method
- Previous methods:
  - Model-free methods
- Model of what?
  - Model of the data

  → Probability density function (PDF)

Two-step Bayesian Decision

- Ideal Condition for Bayesian Decision

  Given the number of classes (states) $\omega_i, i = 1, \cdots, c$, the prior and conditional densities

  $$P(\omega_i),\ P(\boldsymbol{x}|\omega_i),\ i = 1, \cdots, c$$

- Usual situations we face

  Given the number of classes (states) $\omega_i, i = 1, \cdots, c$ and a set of samples in each class $\mathcal{X}_i$

- Two steps:
  - Estimate $P(\omega_i)$ and $p(x|\omega_i)$ from the samples
  - And use the estimated $\hat{P}(\omega_i)$ and $\hat{p}(x|\omega_i)$ in Bayesian Decision

Homework

- Computer exercises (Ex5)
  - Write your own code of Naïve Bayes Classifier and do experiment on the ICU data.
- Deadline:
  - Nov. 17 (Wednesday), 23:00

D


See you next week
for
Probability Density Estimation & HMM
Xuegong Zhang
61