Foundations of Machine Learning
CentraleSupélec — Fall 2017

5. Bayesian decision theory

Chloé-Agathe Azencott
Centre for Computational Biology, Mines ParisTech
chloe-agathe.azencott@mines-paristech.fr

Learning objectives

After this lecture, you should be able to

- **Apply Bayes rule** for simple inference and decision problems;
- Explain the connection between **Bayes decision rule**, **empirical risk minimization**, **maximum a priori** and **maximum likelihood;**
- Use a graph to express **conditional independence** among random variables;
- Apply the **Naive Bayes** algorithm.

Probability and inference

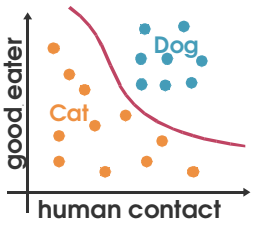
- Result of **tossing a coin:** x in {heads, tails}
  - x = f(**z**) z: **unobserved variables**
  - Replace f(z) (maybe deterministic but unknown) with the **random variable** X in {0, 1} drawn from a **probability distribution** P(X=x).

- We do not know P but a **sample** $X = \{x^i\}_{i=1,\dots,n}$
- Goal: **approximate P** (from which X is drawn)

- **Prediction** of next toss:

Classification

- Cat vs. dog
  - Cat = 1 (positive)
  - Dog = 0 (negative)
  - x_1 = human contact
  - x_2 = good eater
- **Prediction:**



$$\hat{y} = \begin{cases} 1 & \text{if} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y} = \begin{cases} 1 & \text{if} \\ 0 & \text{otherwise} \end{cases}$$

Bayes rule

Reverend Thomas Bayes

170?-1761

… possibly

Bayes rule

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

Example: rare disease testing

- test is correct 99% of the time
- disease prevalence = 1 out of 10,000

What is the probability that a patient that tested positive actually has the disease?

99% ? 90% ? 10% ? 1% ?

Bayes rule

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

prior, likelihood, posterior, evidence

$$P(y=0)+P(y=1)=1$$
$$P(y=0|\boldsymbol{x})+P(y=1|\boldsymbol{x})=1$$

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|y=1)P(y=1) + p(\boldsymbol{x}|y=0)P(y=0)$$

Bayes' decision rule:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|\boldsymbol{x}) > P(y=0|\boldsymbol{x}) \\ 0 & \text{otherwise.} \end{cases}$$

Maximum A Posteriori criterion

- **MAP decision rule:**
  - pick the hypothesis that is most probable
  - i.e. **maximize the posterior**

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

$$\Lambda_{\text{MAP}}(\boldsymbol{x}) = \frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})}$$

- **Decision rule:**

Likelihood ratio test (LRT)

$\Lambda_{\text{MAP}}(\boldsymbol{x}) = \frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})}$ $\Lambda_{\text{MAP}}(\boldsymbol{x}) >^{?} 1$ $P(y=c|\boldsymbol{x}) = \frac{P(y=c)p(\boldsymbol{x}|y=c)}{p(\boldsymbol{x})}$

$$\Lambda_{\text{MAP}}(\boldsymbol{x}) = \frac{P(y=1)p(\boldsymbol{x}|y=1)p(\boldsymbol{x})}{P(y=0)p(\boldsymbol{x}|y=0)p(\boldsymbol{x})}$$

p(**x**) does not affect the decision rule.

- **Likelihood ratio test:**

  test whether the **likelihood ratio** Λ(**x**) is larger than

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)}$$

Example: LRT decision rule

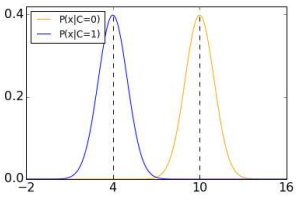
$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} >^{?} \frac{P(y=0)}{P(y=1)}$$

Assuming the likelihoods below <u>and equal priors,</u> derive a decision rule based on the LRT.

$p(x|y=1) \sim \mathcal{N}(4,1)$ $p(x|y=0) \sim \mathcal{N}(10,1)$

$Z \sim \mathcal{N}(\mu, \sigma^2)$:

$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(z-\mu)^2/(2\sigma^2)}$



Maximum likelihood criterion

- Consider **equal priors** P(y=1) = P(y=0)

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} \gtrless^{?} \frac{P(y=0)}{P(y=1)} \quad 1$$

- Bayes decision rule seeks to maximize P(x|y=c) and is hence called the **Maximum Likelihood criterion**

$$\Lambda_{\mathrm{ML}}(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)}$$

  - **Decision rule:**
    If $\Lambda_{ML}(x) > 1$ then choose y=1 else choose y=0

Bayes rule for K > 2

- **Bayes rule:**

$$P(y=c_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=c_k)P(y=C_k)}{\sum_{l=1}^{K} p(\boldsymbol{x}|y=c_l)P(y=c_l)}$$

- $P(y=c_k) \qquad P(y=c_1) + P(y=c_2) + \cdots + P(y=c_K)$
- **What is the decision rule?**

Risk minimization

Losses and risks

- So far we've assumed all errors were **equally costly.**

  But misclassfying a cancer sufferer as a healthy patient is much more problematic than the other way around.

- **Action α_k:** assigining class c_k
- **Loss:** quantify the cost λ_{kl} of taking action α_k when the true class is c_l
- **Expected risk:**

$$R(\alpha_k|\boldsymbol{x}) = \sum_{l=1}^{K} \lambda_{lk} P(y=c_l|\boldsymbol{x})$$

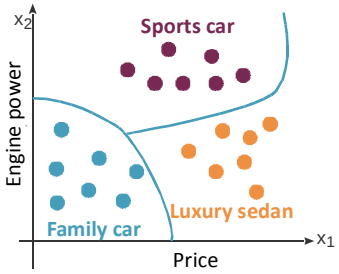
- **Decision (Bayes Classifier):** $\arg\min_k R(\alpha_k|\boldsymbol{x})$

Price

Discriminant functions

Classification = find K **discriminant functions** f_k s.t. **x** is assigned class c_k if k = argmax f_l(**x**)

- Bayes classifier: $f_k(\boldsymbol{x}) = -R(\alpha_k|\boldsymbol{x})$
- Defines K **decision regions** $R_k = \{\boldsymbol{x} : f_k(\boldsymbol{x}) = \max_l f_l(\boldsymbol{x})\}$



0/1 Loss

- All misclassifications are **equally costly.**
- λ_{kl} = 0 if k=l and 1 otherwise

$$\begin{aligned} R(\alpha_k|\boldsymbol{x}) &= \sum_{l=1}^{K} \lambda_{lk} P(y = C_l|\boldsymbol{x}) \\ &= \sum_{l \neq k} P(y = C_l|\boldsymbol{x}) \\ &= 1 - P(y = C_k|\boldsymbol{x}) \end{aligned}$$

- **Minimizing the risk:**
  - choose the most probable class (MAP)
  - this is equivalent to the Bayes decision rule.

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} >^{?} \frac{(\lambda_{10} - \lambda_{00})P(y=0)}{(\lambda_{01} - \lambda_{11})P(y=1)}$$

Maximum likelihood criterion

- Consider **equal priors** P(y=1) = P(y=0)
- Consider the **0/1 loss function**

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} >^{?} \frac{(\lambda_{10} - \lambda_{00})P(y=0)}{(\lambda_{01} - \lambda_{11})P(y=1)}$$

Reject

- Add an artificial "reject" class (K+1) for **refusing to take a decision.**

  E.g. Zip code detection.

- $\lambda_{kl} = \begin{cases} 0 \text{ if } k = k \\ \lambda \text{ if } k = K+1 \\ 1 \text{ otherwise} \end{cases}$

$$R(\alpha_k|\boldsymbol{x}) = \sum_{l \neq k} P(y = c_l|\boldsymbol{x}) = 1 - P(y = c_k|\boldsymbol{x})$$

$$R(\alpha_{K+1}|\boldsymbol{x}) = \sum_{l=1}^{K} \lambda P(y = c_l|\boldsymbol{x}) = \lambda$$

- **Decision:**

$$\hat{y} = c_k \text{ if } P(y = c_k|\boldsymbol{x}) > P(y = c_l|\boldsymbol{x}) \text{ for all } l \neq k \text{ and } P(y = c_k|\boldsymbol{x}) > 1 - \lambda$$

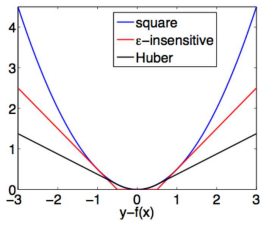
  else reject.

  Only meaningful if $0 < \lambda < 1$

Losses for regression

- **Square loss:** $L(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$
- **ε-insensitive loss:** $L(f(\mathbf{x}), y) = (|f(\mathbf{x}) - y| - \varepsilon)_+$
- **Huber loss:** mix of linear and quadratic

$$L_\delta(f(\boldsymbol{x}), y) = \begin{cases} \frac{1}{2}(y - f(\boldsymbol{x}))^2 & \text{if } |y - f(\boldsymbol{x})| \leq \delta \\ \delta|y - f(\boldsymbol{x})| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$



Empirical risk minimization (ERM)

- **Loss:** $L(f(\mathbf{x}), y)$ small when $f(\mathbf{x})$ predicts y well
- **Expected risk:**

$$R = \mathbb{E}[L(f(\boldsymbol{x}), y)]$$

- **Empirical risk:**

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(f(\boldsymbol{x}^i), y^i)$$

- The **ERM estimator** of the functional class F is the solution, when it exists, of:

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} R_n(f)$$

Solving ERM

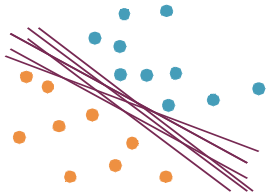
- There can sometimes be an **explicit analytical solution**
- Otherwise: **convex optimization** (if the loss function is convex in f)
- **Limits of ERM:**
  - **ill-posed**
  - **not statistically consistent**

    This is particularly true in **high dimension.**

ERM is ill-posed

- **Well-posed problems** (Hadamard):

  Mathematical models of physical phenomena such that
  - a solution exists;
  - the solution is unique;
  - the solution's behavior changes continuously with the initial conditions.
- It can be that **an infinite number of solutions minimize the empirical risk** to zero.



ERM is not statistically consistent

- **Statistical consistency:** Estimator θ_N of θ that converges in probability towards θ as N increases.

$$\forall \epsilon > 0 \quad \lim_{N \to \infty} Pr(|\theta_N - \theta| \geq \epsilon) = 0$$

- From the **law of large numbers**,

$$\forall f \in \mathcal{F}, \quad R_N(f) \xrightarrow[N \to \infty]{} R(f)$$

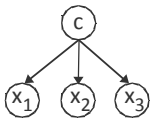
  but this isn't enough to guarantee that minimizing $R_N(f)$ gives a good estimator of the minimizer of R(f).
- Vapnik showed that this is only true if the capacity of hypothesis space $\mathcal{F}$ is "not too large".

Multivariate classification: Naive Bayes

- Multivariate classification: **x** is multidimensional
- Assume the variables x_1, x_2, … x_p are **conditionally independent**: $p(x_{j_1}|y=c, x_{j_2}) = p(x_{j_1}|y=c)$

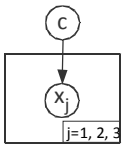
Graphical representation

- We can use a graph to represent **conditional independence**:
  - arc from c to x_j means the distribution of X_j **depends** on c
  - no arc between X_{j1} and X_{j2} means that X_{j1} and X_{j2} are **independent given C**:
    $p(x_{j_1}|y=c, x_{j_2}) = p(x_{j_1}|y=c)$



- A **plate** represents repeated structure:

  all X_j inside the same plate follow the same probability distribution.



Naive Bayes

- Multivariate classification: x is multidimensional
- Assume the variables x_1, x_2, … x_p are **conditionally independent**: $p(x_{j_1}|y=c, x_{j_2}) = p(x_{j_1}|y=c)$

$$P(y=c_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=c_k)P(y=c_k)}{\sum_{l=1}^{K} p(\boldsymbol{x}|y=c_l)P(y=c_l)}$$

$$p(x_1,\ldots,x_p|y=c) = p(x_1|y=c)p(x_2|y=c)\ldots p(x_p|y=c)$$

Hence:

$$P(y=c|x_1,\ldots,x_p) = \frac{1}{Z}P(y=c)p(x_1|y=c)p(x_2|y=c)\ldots p(x_p|y=c)$$

scaling factor, independent of c_k

Maximum a posteriori estimation

- **MAP decision rule:** pick the hypothesis that is most probable
- For Naive Bayes:

$$\hat{y} = \arg \max_{k=1,\ldots,K} p(y = c_k) \prod_{i=1}^{n} p(\boldsymbol{x}^i | y = c_k)$$

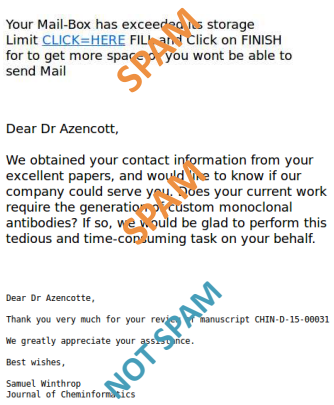
Naive Bayes spam filtering

- Input: email

  bag of words

  $(x_1, x_2, \ldots x_p) = (0, 1, \ldots, 0)$

  rich CLICK viagra



- Output: spam / ham
- Naive Bayes assumption:

  conditional independence

$$P(y = c_k | \boldsymbol{x}) = \frac{p(\boldsymbol{x} | y = c_k) P(y = c_k)}{\sum_{l=1}^{K} p(\boldsymbol{x} | y = c_l) P(y = c_l)}$$

- **P(spam|(x_1, x_2, … x_p))**

  = 1/Z p(spam) p(x_1|spam) p(x_2|spam) … p(x_p|spam)

- **P(ham|(x_1, x_2, … x_p))**

  = 1/Z p(ham) p(x_1|ham) p(x_2|ham) … p(x_p|ham)

- **Decision:**

  If P(spam|(x_1, x_2, …, x_p)) > P(ham|(x_1, x_2, …, x_p)) then spam else ham

- **Inference:** we need to determine

  p(spam), p(ham), p(x_j|spam), p(x_j|ham)

- **Bernouilli Naive Bayes:**
  - Each email is the outcome of p Bernouilli trials
  - **Naive assumption:** the trials are independent

    word co-occurences in a category aren't independent

    still, independence assumptions can give good results

$$p(x_j|\text{spam}) = {p_j}^{x_j}(1-p_j)^{(1-x_j)}$$

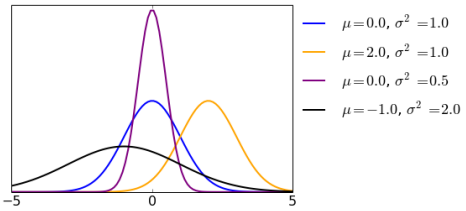
  - **Direct estimate of p_j:** $p_j = S_j / S$
    - S = # spams in train set
    - S_j = # spams containing word j in train set

Gaussian naive Bayes

- Assume

  $p(x_j|y=c_k)$ **univariate Gaussian**

$$p(x_j|y=c_k) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_j-\mu)^2/(2\sigma^2)}$$



Bayesian model selection

- **Priors on model:** p(model)

$$p(\text{model}|\text{data}) = \frac{p(\text{data}|\text{model})p(\text{model})}{p(\text{data})}$$

- Regularization ≡ prior that favors simpler models.
- Take the log

$$\log p(\text{model}|\text{data}) = \log p(\text{data}|\text{model}) + \log p(\text{model}) - c$$

≡ training error ≡ model complexity

- MAP similar to minimizing

  E' = empirical error + **λ** model complexity

Summary

$$\underbrace{P(y=c|\boldsymbol{x})}_{\text{posterior}} = \frac{\overbrace{P(y=c)}^{\text{prior}}\overbrace{p(\boldsymbol{x}|y=c)}^{\text{likelihood}}}{\underbrace{p(\boldsymbol{x})}_{\text{evidence}}}$$

- **Bayes decision rule ≡ likelihood ratio test**

  choose the most probable class, given evidence (data) and prior belief.

- Equivalent to **minimizing Bayes risk**

  usually achieved approximately through **empirical risk minimization** (not equivalent!!)

- For the 0/1 loss, equivalent to **maximizing the posterior.**

- For the 0/1 loss and equal priors (uniform prior), equivalent to **maximizing the likelihood.**

References

- *A Course in Machine Learning.*
  `http://ciml.info/dl/v0_99/ciml-v0_99-all.pdf`
  - **Bayes classifier:** Chap 2.1
  - **LRT:** Chap 9.4
  - **Naive Bayes:** Chap 9.3
- *The Elements of Statistical Learning.*
  `http://web.stanford.edu/~hastie/ElemStatLearn/`
  - **Bayes classifier:** Chap 2.4
  - **Maximum Likelihood:** Chap 2.6.3, Chap 8.3
- **Probabilistic machine learning**
  `https://www.repository.cam.ac.uk/bitstream/handle/1810/248538/Ghahramani%202015%20Nature`
- **Spam detection:** `http://www.paulgraham.com/spam.html`
- **Naive Bayes:**
  `https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf`