Bayesian decision theory

Introduction to Machine Learning – GIF-7015

Professor: Christian Gagné

Week 2

UNIVERSITÉ LAVAL

2.1 Bayes formula

Review of basic statistical concepts

- Random experiment ($\mathcal{E}$): an experiment for which the outcome cannot be predicted in advance with certainty
- Sample space (U): the set of all possible outcomes or results of an experiment
  - Discrete sample space: finite set of possible outcomes
  - Continuous sample space: the possible outcomes are not enumerable
- Random event (A): result of a random experiment, subset of the sample space ($A \subset U$)
- Probability ($P(A)$): associate a real number representing the application of a given event (A) related to a random experiment ($A \subset U$), satisfying the axioms of probabilities
  1. $0 \leq P(A) \leq 1, \forall A$
  2. $P(U) = 1$
  3. Suppose the events A_i, $i = 1, \ldots, n$ are mutually exclusive ($A_i \cap A_j = \emptyset, \forall j \neq i$), then $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$

Probability and inference

- Tossing a coin: $U = \{\text{tail}, \text{head}\}$
- Random variable $X = \{0, 1\}$ (0=head, 1=tail)
  - Bernoulli distribution: $P(x \in X) = (1 - p_1)^{1-x} \, p_1^x$
- Set of samples $\mathbf{X}$ drawn according to a probability distribution parameterized by p_1 (tail probability)
  - Set of N samples: $\mathbf{X} = \{x^t\}_{t=1}^N$ with $x^t \in X$
  - Estimate of p_1 by sampling: $\hat{p}_1 = \frac{\#tails}{\#tosses} = \frac{\sum_{i=1}^N x^t}{N}$
- Prediction of the next toss x^{N+1}: if $\hat{p}_1 > 0.5$ then tail, otherwise head
- Example of outcomes: $\mathbf{X} = \{1, 1, 1, 0, 1, 0, 0, 1, 1\}$
  - Estimation of the probability: $\hat{p}_1 = \frac{\sum_{t=1}^N x^t}{N} = \frac{6}{9}$

Classification

- Example of credit risk assessment
  - Input data: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, with x_1 as income and x_2 the amount of savings
  - Possible classes: $C \in \{0, 1\}$ where $C = 1$ denotes an individual at high risk of default and $C = 0$ an individual at low risk of default
- If we know $P(C|x_1, x_2)$ then:
  - Assign: $\begin{cases} C = 1 & \text{if } P(C = 1|x_1, x_2) > 0.5 \\ C = 0 & \text{otherwise} \end{cases}$
- Equivalent formulation:
  - Assign: $\begin{cases} C = 1 & \text{if } P(C = 1|x_1, x_2) > P(C = 0|x_1, x_2) \\ C = 0 & \text{otherwise} \end{cases}$

Conditional probability

- Conditional probability $P(E|F)$: probability that the event E will occur if the event F has occurred:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

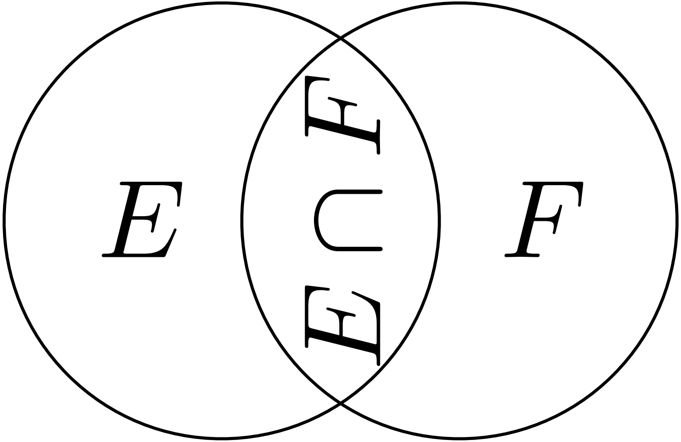
- Since $\cap$ is commutative:

$$P(E \cap F) = P(E|F)\,P(F) = P(F|E)\,P(E)$$

- Bayes formula:

$$P(F|E) = \frac{P(E|F)\,P(F)}{P(E)}$$

Venn diagram and Bayes formula



$$P(E \cap F) = P(E|F)\,P(F) = P(F|E)\,P(E) = P(F \cap E)$$

Bayes formula

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

- Prior probability ($P(C)$): probability of observing an instance of the class C
- Class likelihood ($p(\mathbf{x}|C)$): likelihood that an observation of the class C is $\mathbf{x}$
- Evidence ($p(\mathbf{x})$): likelihood of observing the data $\mathbf{x}$
- Posterior probability ($P(C|\mathbf{x})$): probability that an observation $\mathbf{x}$ belongs to the class C

Bayes formula

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

- Sum of prior probabilities: $P(C=0) + P(C=1) = 1$
- Sum of posterior probabilities: $P(C=0|\mathbf{x}) + P(C=1|\mathbf{x}) = 1$
- Evidence: $p(\mathbf{x}) = P(C=1)\,p(\mathbf{x}|C=1) + P(C=0)\,p(\mathbf{x}|C=0)$

Example: Bayes formula

- Vehicle observation
  - Probability of observing a car, $P(C = 1) = 0.7$
  - Probability of observing another vehicle, $P(C = 0) = 0.3$
- A given vehicle observation $\mathbf{x}$
  - Likelihoods of the observation: $p(\mathbf{x}|C = 1) = 1.1$, $p(\mathbf{x}|C = 0) = 0.4$
- Evidence

$$\begin{aligned} p(\mathbf{x}) &= p(\mathbf{x}|C = 1)\,P(C = 1) + p(\mathbf{x}|C = 0)\,P(C = 0) \\ &= 1.1 \cdot 0.7 + 0.4 \cdot 0.3 = 0.77 + 0.12 = 0.89 \end{aligned}$$

- Posterior probabilities

$$\begin{aligned} P(C = 1|\mathbf{x}) &= \frac{P(C = 1)\,p(\mathbf{x}|C = 1)}{p(\mathbf{x})} = \frac{0.7 \cdot 1.1}{0.89} = \frac{0.77}{0.89} = 0.865 \\ P(C = 0|\mathbf{x}) &= \frac{P(C = 0)\,p(\mathbf{x}|C = 0)}{p(\mathbf{x})} = \frac{0.3 \cdot 0.4}{0.89} = \frac{0.12}{0.89} = 0.134 \end{aligned}$$

2.2 Bayesian Decision Theory

Bayes formula with several classes

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

- $P(C_i) \geq 0$ et $\sum_{i=1}^{K} P(C_i) = 1$
- Choose class C_i for data $\mathbf{x}$ according to $C_i = \operatorname*{argmax}_{k=1}^{K} P(C_k|\mathbf{x})$

Loss function

- Not all decisions have the same impact
  - Lending money to a high-risk client versus not lending to a low-risk client
  - Medical diagnosis: possible impacts of not detecting a serious illness
  - Intrusion detection
- Quantify with a loss function $\mathcal{L}(\alpha_i, C_j)$
  - Perform an action α_i while the actual class is C_j

Risk

- Expected risk of an action α:

$$R(\alpha|\mathbf{x}) = \sum_{k=1}^{K} \mathcal{L}(\alpha, C_k)\, P(C_k|\mathbf{x})$$

- Action minimizing risk:

$$\alpha^* = \underset{\forall \alpha}{\text{argmin}}\, R(\alpha|\mathbf{x})$$

- Modifying the loss function changes the risk
    - Modifying the cost associated with a false negative relative to the cost of a false positive

Confusion matrix (two classes)

		Decision	
		α_0	α_1
Actual	C_0	0	λ_{FP}
	C_1	λ_{FN}	0

- $\mathcal{L}(\alpha = 1, C = 0) = \lambda_{\mathrm{FP}}$: cost of a false positive
- $\mathcal{L}(\alpha = 0, C = 1) = \lambda_{\mathrm{FN}}$: cost of a false negative

Confusion matrix (K classes)

	α_0	α_1	$\cdots$	α_K
C_0	0	$\lambda_{1,0}$	$\cdots$	$\lambda_{K,0}$
C_1	$\lambda_{0,1}$	0	$\cdots$	$\lambda_{K,1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_K	$\lambda_{0,K}$	$\lambda_{1,K}$	$\cdots$	0

Zero-one loss function

- Zero-one loss function:

$$\mathcal{L}(\alpha_i, C_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

- Corresponding risk:

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{k=1}^{K} \mathcal{L}(\alpha_i, C_k)\, P(C_k|\mathbf{x}) \\ &= \sum_{k \neq i} P(C_k|\mathbf{x}) \\ &= 1 - P(C_i|\mathbf{x}) \end{aligned}$$

- Optimal decision:

$$\alpha^* = \underset{\alpha_k = \alpha_1}{\overset{\alpha_K}{\operatorname{argmax}}}\, P(C_k|\mathbf{x})$$

Reject option

- For many applications, a bad classification can have a huge impact
  - Addition of a reject option in case of doubt, action α_{K+1}
- Zero-one loss function with reject option:

$$\mathcal{L}(\alpha_i, C_j) = \begin{cases} 0 & \text{if } i = j \\ \lambda & \text{if } i = K+1 \\ 1 & \text{otherwise} \end{cases}$$

- In that case:

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{k \neq i} P(C_k|\mathbf{x}) = 1 - P(C_i|\mathbf{x}) \\ R(\alpha_{K+1}|\mathbf{x}) &= \sum_{k=1}^{K} \lambda P(C_k|\mathbf{x}) = \lambda \end{aligned}$$

Optimal decision with reject option

- Optimal decision with reject option:

$$\alpha^* = \underset{\alpha_k=\alpha_1}{\overset{\alpha_{K+1}}{\text{argmin}}} R(\alpha_k|\mathbf{x})$$

- Optimal decision for zero-one loss function with reject option:

$$\alpha^* = \begin{cases} \alpha_{K+1} & \text{if } P(C_j|\mathbf{x}) < 1-\lambda, \forall j = 1, \ldots, K \\ \underset{\alpha_j=\alpha_1}{\overset{\alpha_K}{\text{argmax}}} P(C_j|\mathbf{x}) & \text{otherwise} \end{cases}$$

Confusion matrix (K classes and reject option)

	α_0	α_1	$\cdots$	α_K	α_{K+1}
C_0	0	$\lambda_{1,0}$	$\cdots$	$\lambda_{K,0}$	$\lambda_{K+1,0}$
C_1	$\lambda_{0,1}$	0	$\cdots$	$\lambda_{K,1}$	$\lambda_{K+1,1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_K	$\lambda_{0,K}$	$\lambda_{1,K}$	$\cdots$	0	$\lambda_{K+1,K}$

Discriminant function

- Discriminant functions for classification: $\alpha^t = \underset{\alpha_i=\alpha_1}{\overset{\alpha_K}{\operatorname{argmax}}}\, \mathrm{h}_i(\mathbf{x}^t)$
  - In the Bayesian case (general): $\mathrm{h}_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$
  - Bayesian with zero-one loss function: $\mathrm{h}_i(\mathbf{x}) = P(C_i|\mathbf{x})$
  - Ignoring normalization relative to $p(\mathbf{x})$: $\mathrm{h}_i(\mathbf{x}) = p(\mathbf{x}|C_i)\,P(C_i)$
- *Decision regions*: division of the input space into K regions:
  - $\mathcal{R}_1, \ldots, \mathcal{R}_K$ où $\mathcal{R}_i = \{\mathbf{x}|\mathrm{h}_i(\mathbf{x}) = \max_{\forall k} \mathrm{h}_k(\mathbf{x})\}$
- Decision regions are separated by *decision boundaries*
- Two-class case is a *dichotomizer*, $K \geq 3$ classes is a *plurichotomizer*

Regions and decision boundaries