Introduction to Artificial Intelligence (INFO8006)

Exercises 5 – Learning

November 10, 2023

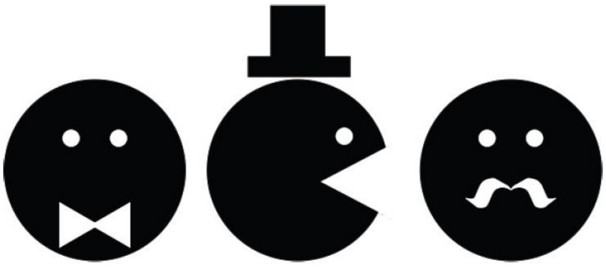
Learning outcomes

At the end of this session you should be able to

- define and apply maximum a posteriori (MAP) estimation;
- define and apply maximum likelihood estimation (MLE);
- define and apply linear regression.

Exercise 1 Pacbaby (UC Berkeley CS188, Spring 2014)

Pacman and Pacwoman have been searching for each other in the maze. Pacwoman has been pregnant with a baby, and just this morning she has given birth to Pacbaby[1]. Because Pacbaby was born before Pacman and Pacwoman were reunited in the maze, he has never met his father. Naturally, Pacwoman wants to teach Pacbaby to recognize his father, using a set of pictures of Pacman. She also has several pictures of ghosts to use as negative examples.

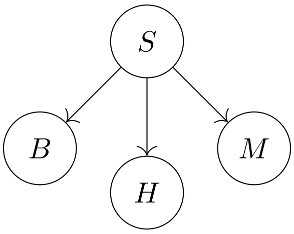


Because the pictures are black and white, and were taken from various angles, Pacwoman has decided to teach Pacbaby to identify Pacman based on salient features: the presence of a bowtie B, hat H or mustache M. The following table summarizes the content of the pictures. Each feature takes realization in $\{0, 1\}$, where 0 and 1 mean the feature is respectively absent and present. The subject of the picture is described by a random variable $S \in \{0, 1\}$, where 0 is a ghost and 1 is Pacman.

B	H	M	S
0	0	0	1
1	0	0	0
1	0	1	1
1	1	0	0
0	1	0	0
1	1	1	1

1. Suppose Pacbaby has a Naive Bayes based brain. Draw the Bayesian network that would represent the dependencies between S, B, H and M for Pacbaby.

[1] Congratulations!



2. Write the Bayesian classification rule for this problem, *i.e.* the formula that given a data point (b, h, m) returns the most likely subject. Write the formula in terms of conditional and prior probabilities. What does the formula become under the assumptions of Pacbaby ?

   Given (b, h, m), the most likely subject is given by the *maximum a posteriori* (MAP) estimation

$$\begin{aligned} s_{\text{MAP}} &= \arg\max_s P(s|b, h, m) \\ &= \arg\max_s P(b, h, m|s)P(s). \end{aligned}$$

   Under the naive Bayes assumptions of Pacbaby, B, H and M become independent conditionally to S, *i.e.* $P(B, H, M|S) = P(B|S)P(H|S)P(M|S)$. Then, the formula becomes

$$s_{\text{MAP}} = \arg\max_s P(b|s)P(h|s)P(m|s)P(s).$$

3. What are the parameters of this model? Give estimates of these parameters according to the pictures provided by Pacwoman.

   The parameters of the model are the elements of the prior vector $P(S)$ and the (conditional) probability matrices $P(B|S)$, $P(H|S)$ and $P(M|S)$. An (unbiased) estimation of these elements can be computed as the frequency of their respective events within the learning set (of pictures).

S	$P(S)$	$P(B = 1\|S)$	$P(H = 1\|S)$	$P(M = 1\|S)$
0	$\frac{3}{6}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{0}{3}$
1	$\frac{3}{6}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$

4. Pacman eventually shows up wearing a bowtie, but no hat or mustache. Will Pacbaby recognize his father?

   Pacbaby will recognize his father if $s_{\text{MAP}} = 1$ for $(b, h, m) = (1, 0, 0)$. Using the parameters estimated previously, we have

$$\begin{aligned} P(b|0)P(h|0)P(m|0)P(0) &= \frac{2}{3} \times \left(1 - \frac{2}{3}\right) \times \left(1 - \frac{0}{3}\right) \times \frac{3}{6} \approx 0.111 \\ P(b|1)P(h|1)P(m|1)P(1) &= \frac{2}{3} \times \left(1 - \frac{1}{3}\right) \times \left(1 - \frac{2}{3}\right) \times \frac{3}{6} \approx 0.074. \end{aligned}$$

   Therefore, $s_{\text{MAP}} = 0$, meaning that Pacbaby will *not* recognize his father.

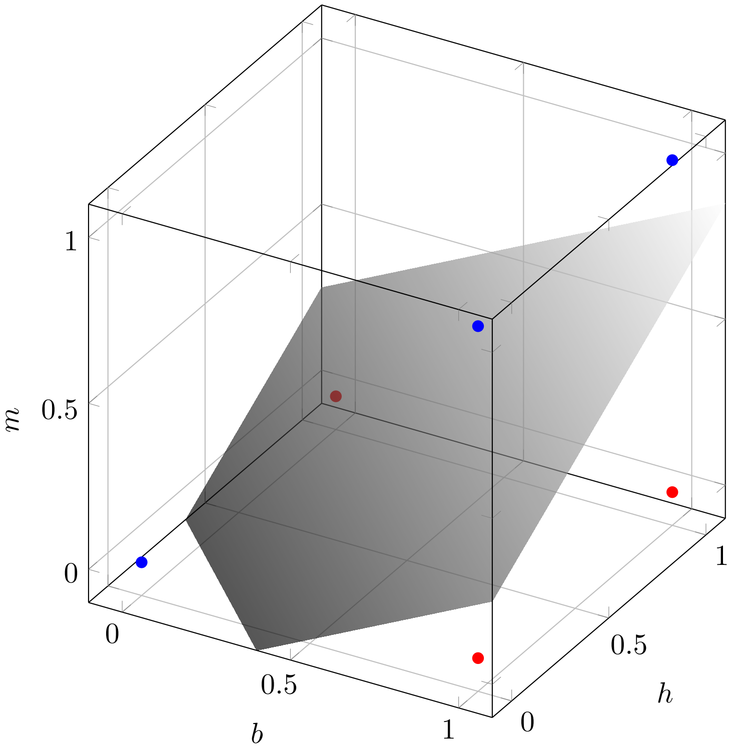
5. If Pacbaby had a perceptron based brain, meaning that he is limited to learn linear classification rules, would he be able to learn a rule that makes no mistakes on the set of pictures? In other words, is the learning set *linearly separable*?

For a training set $\{(x_i, y_i)\}$ with $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$, a linear separation consists in *any* hyperplane parameters $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

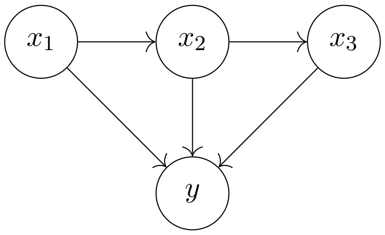
$$y_i = \text{sign}(w^T x_i + b)$$

for all (x_i, y_i). There exists several ways of fitting these parameters to a learning set, but it is very tedious by hand. However, in low dimension (2D, 3D), it is possible to solve this problem visually. We can draw the data points on a grid with different colors for each class and try to find a plane that separates best the two classes.

In our case, we draw the $S = 0$ class in red and the $S = 1$ class in blue, and see that there is a plane that separates them perfectly. Hence, the learning set is linearly separable.



Exercise 2 Predict your grade



The hereabove Bayesian network represents how the final grade of a class is computed. In this model, x_1, x_2 and x_3 respectively denote the grades obtained by a student at the homework, project and exam. The teaching assistant that grades the homework also grades the project and the exam, which introduces a slight bias in the corrections. In particular, $x_2 \sim \mathcal{N}(a_1 x_1 + \mu_2, \sigma_2^2)$ and $x_3 \sim \mathcal{N}(a_2 x_2 + \mu_3, \sigma_3^2)$. Finally, $y \sim \mathcal{N}(a_3 x_1 + a_4 x_2 + a_5 x_3 + \mu_y, \sigma_y^2)$ stands for the final grade, which is a linear combination of the grades obtained by the student during the semester plus some Gaussian noise due to rounding errors. Answer the following questions about this model.

1. Assuming the parameters of the model are known, what is the expected value of y given x_1 and x_2.

   Our task is to find the expectation

   $$\mathop{\mathbb{E}}_{p(y|x_1,x_2)}[y] = \int y \, p(y|x_1, x_2) \, \mathrm{d}y.$$

   We know that

   $$p(y|x_1, x_2) = \int p(y|x_1, x_2, x_3) \, p(x_3|x_1, x_2) \, \mathrm{d}x_3,$$

   where $p(y|x_1, x_2, x_3)$ and $p(x_3|x_1, x_2)$ are linear Gaussian distributions given in the statement. Therefore, we have

   $$p(y|x_1, x_2) = \mathcal{N}\Big(a_3 x_1 + a_4 x_2 + a_5(a_2 x_2 + \mu_3) + \mu_y, (a_5 \sigma_3)^2 + \sigma_y^2\Big)$$

   and, by definition of a Gaussian distribution,

   $$\mathop{\mathbb{E}}_{p(y|x_1,x_2)}[y] = a_3 x_1 + (a_4 + a_5 a_2) x_2 + a_5 \mu_3 + \mu_y.$$

2. Suppose now that the model's parameters are unknown. Given a learning set $d = \{(x_{i,1}, x_{i,2}, y_i)\}$ of N independent and identically distributed points, determine the model that best describes d.

   We know that the distribution of y given x_1 and x_2 takes the form $\mathcal{N}(w_1 x_1 + w_2 x_2 + b, \sigma^2)$. Then, our task is to find the parameters $h = (w_1, w_2, b, \sigma)$ that maximize the likelihood of d, *i.e.* the *maximum likelihood estimation* (MLE)

   $$\begin{aligned} h_{\text{MLE}} &= \arg\max_w p(d|h) \\ &= \arg\max_h \prod_i p(x_i, y_i|h) \\ &= \arg\max_h \log \prod_i p(x_i, y_i|h) \\ &= \arg\max_h \sum_i \log p(x_i, y_i|h) \end{aligned}$$

$$
\begin{aligned}
&= \arg\max_h \sum_i \log p(y_i|h, x_i) + \log p(x_i) \\
&= \arg\max_h \sum_i \log p(y_i|h, x_i) \\
&= \arg\max_h \sum_i \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w^T x_i - y_i)^2}{2\sigma^2}\right)\right] \\
&= \arg\max_h \sum_i -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(w^T x_i - y_i)^2}{2\sigma^2} \\
&= \arg\min_h \log\sigma^2 + \frac{1}{\sigma^2}\frac{1}{N}\sum_i (w^T x_i - y_i)^2,
\end{aligned}
$$

where $x_i = \begin{pmatrix} x_{i,1} & x_{i,2} & 1 \end{pmatrix}^T$ and $w = \begin{pmatrix} w_1 & w_2 & b \end{pmatrix}^T$. In the last expression, we observe that the summation term is independent from σ. Therefore,

$$
w_{\text{MLE}} = \arg\min_w \sum_i (w^T x_i - y_i)^2,
$$

which exactly corresponds to a *linear regression* problem. Then, we find w_{MLE} by canceling the gradient with respect to w, *i.e.*

$$
\begin{aligned}
0 &= \nabla_w \sum_i (w^T x_i - y_i)^2 \\
&= \nabla_w \sum_i (w^T x_i - y_i)(w^T x_i - y_i) \\
&= \nabla_w \sum_i (w^T x_i)^2 + y_i^2 - 2w^T x_i y_i \\
&= \nabla_w \left(w^T X^T X w + Y^T Y - 2w^T X^T Y\right) \\
&= 2X^T X w - 2X^T Y
\end{aligned}
$$

where $X = (x_i^T) \in \mathbb{R}^{N\times 3}$ and $Y = (y_i) \in \mathbb{R}^N$. Finally, we have

$$
\begin{aligned}
& 0 = X^T X w_{\text{MLE}} - X^T Y \\
\Leftrightarrow \quad & w_{\text{MLE}} = (X^T X)^{-1} X^T Y.
\end{aligned}
$$

Afterwards, we find σ_{MLE} such that

$$
\begin{aligned}
\sigma_{\text{MLE}} &= \arg\min_\sigma \log\sigma^2 + \frac{\text{MSE}}{\sigma^2} \\
&= \sqrt{\text{MSE}},
\end{aligned}
$$

where MSE denotes the *mean squared error*

$$
\frac{1}{N}\sum_i ({w_{\text{MLE}}}^T x_i - y_i)^2.
$$

Exercise 3 Heteroscedastic linear regression

What becomes the expression of the weight vector w in the solution of question 2.2 if the noise is different for each sample? In particular, $y_i \sim N(w^T x, \sigma_i^2)$ and we know the values σ_i.

Exercise 4 Ridge regression

One can generalize the linear regression problem to the minimization problem

$$w^* = \arg\min_w \sum_i \ell(w^T x_i - y_i),$$

where ℓ is a *loss* function. Show that $\ell(x) = |x|$ corresponds to assuming the noise follows a Laplace distribution in contrast to $\ell(x) = x^2$, which corresponds to assuming Gaussian noise.

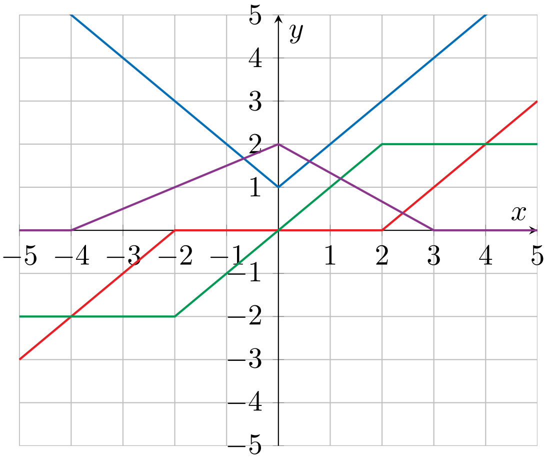
Exercise 5 Learning to play Pacman (August 2020)

You observe a Grandmaster agent playing Pacman. How can you use the moves you observe to train your own agent?

1. Describe formally the data you would collect, the inference problem you would consider, and how you would solve it.
2. How would you design a neural network to control your agent? Define mathematically the neural network architecture, its inputs, its outputs, its parameters, as well as the loss you would use to train it.
3. Discuss the expected performance of the resulting agent when (a) the Grandmaster agent is optimal, and (b) the Grandmaster agent is suboptimal.

Exercise 6 ReLU

For each of the piecewise-linear functions below, write a function $y = f(x)$ as a composition of sums $(+, -)$, ReLU ($\mathrm{ReLU}(x) = \max(x, 0)$) non-linearities, and real parameters (weights and biases) that matches exactly the function over $\mathbb{R}$.



For example, $y = \mathrm{ReLU}(x + 2) - \mathrm{ReLU}(-2x)$ is a valid function.

1. $y = \mathrm{ReLU}(x) - \mathrm{ReLU}(-x) + 1$
2. $y = \mathrm{ReLU}(x - 2) - \mathrm{ReLU}(-x - 2)$
3. $y = \mathrm{ReLU}(-\mathrm{ReLU}(-x + 2) + 4) - 2$
4. $y = \mathrm{ReLU}(-\mathrm{ReLU}(\frac{2}{3}x) - \mathrm{ReLU}(-\frac{1}{2}x) + 2)$

Exercise 7 Escape game (January 2022)

A new virtual escape game came out, and you decide to play it. You arrive in a 5×5 grid world where each cell (x, y) is a room with doors leading to the adjacent rooms. The game's goal is to reach the exit room as fast as possible, but its position is unknown. Furthermore, some regions of the world are full of riddles, and crossing rooms in these regions takes longer. Fortunately, a leaderboard with the players' best times is provided, starting from a few different rooms. Due to rounding errors, you assume that the best times reported in the leaderboard are measurements affected by additive Gaussian noise $\mathcal{N}(0, 1)$.

i	Starting room	Measured best time
1	$(4, 5)$	2.0
2	$(5, 3)$	3.5
3	$(3, 3)$	4.5
4	$(4, 1)$	7.0
5	$(1, 2)$	8.5

From the leaderboard, you wish to learn a heuristic approximating the best time to get to the exit, starting from room (x, y). You decide to use a small neural network as approximator, described by the following parametric function,

$$\begin{aligned} h(x, y; \phi) &= \text{ReLU}(xw_1 + yw_2 + w_3) + \text{ReLU}(xw_4 + yw_5 + w_6) \\ \text{ReLU}(x) &= \max(x, 0), \end{aligned}$$

where $\phi = (w_1, w_2, w_3, w_4, w_5, w_6)$ is the set of parameters/weights of the neural network.

1. Among the following sets of parameters (A, B or C), which one would you use? Justify your answer.

Set	w_1	w_2	w_3	w_4	w_5	w_6
ϕ_A	-1.5	1	4	1	-1.5	6
ϕ_B	-1	1.5	3	0	-1	4
ϕ_C	-2	0.5	4.5	1.5	0	5

A set of parameters is better than another if it "explains" the data better, that is the likeliness of the data given the set of parameter is higher. Our task is to find the set of parameters that maximizes the likelihood of data under the assumed probability model, or maximum likelihood estimation (MLE). In our case, the data d is the leaderboard and consists of independent position-time tuples (x_i, y_i, t_i). From the statement, we gather that the (best) time t is a function of the starting position (x, y), approximated by a neural network $h(x, y; \phi)$. We also learn that our time measurements are affected by Gaussian noise.

$$\begin{aligned} \phi_{\text{MLE}} &= \arg\max_{\phi} p(d|\phi) \\ &= \arg\max_{\phi} \prod_{i=1}^{5} p(x_i, y_i, t_i|\phi) \\ &= \arg\max_{\phi} \prod_{i=1}^{5} p(t_i|x_i, y_i, \phi)\, p(x_i, y_i) \end{aligned}$$

$$
\begin{aligned}
&= \arg\max_{\phi} \prod_{i=1}^{5} \mathcal{N}(t; h(x, y; \phi), 1) \\
&= \arg\max_{\phi} \sum_{i=1}^{5} \log\left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(t_i - h(x_i, y_i; \phi))^2}{2}\right)\right] \\
&= \arg\min_{\phi} \sum_{i=1}^{5} (t_i - h(x_i, y_i; \phi))^2
\end{aligned}
$$

Now that we have an objective $L(\phi)$ to minimize, we can evaluate it for each of the proposed sets and select the best one.

$$
\begin{aligned}
L(\phi_A) &= \sum_{i=1}^{5} (t_i - h(x_i, y_i; \phi_A))^2 = (2.0 - 5.5)^2 + (3.5 - 6.5)^2 + \cdots = 29.75 \\
L(\phi_B) &= 35.75 \\
L(\phi_C) &= 205.25
\end{aligned}
$$

The best set is ϕ_A.

2. You now assume a Gaussian prior $\mathcal{N}(0, 1)$ on each parameter. Which set of parameters in the table above would you now choose? Justify your answer.

   Now that we hold a prior belief on the parameters, our task is to find the most likely set of parameters given the data, that is maximum a posteriori (MAP) estimation.

$$
\begin{aligned}
\phi_{\text{MAP}} &= \arg\max_{\phi} p(\phi|d) \\
&= \arg\max_{\phi} p(d|\phi)p(\phi) \\
&= \arg\max_{\phi} \prod_{i=1}^{5} p(x_i, y_i, t_i|\phi) \prod_{j=1}^{6} p(w_j) \\
&= \arg\max_{\phi} \prod_{i=1}^{5} \mathcal{N}(t; h(x, y; \phi), 1) \prod_{j=1}^{6} \mathcal{N}(w_j; 0, 1) \\
&= \arg\min_{\phi} \sum_{i=1}^{5} (t_i - h(x_i, y_i; \phi))^2 + \underbrace{\sum_{j=1}^{6} (w_j)^2}_{\|\phi\|^2}
\end{aligned}
$$

   Here again, we have an objective $L'(\phi)$ to minimize, which we evaluate for each of the proposed sets.

$$
\begin{aligned}
L'(\phi_A) &= L(\phi_A) + \sum_{j=1}^{6} (w_j)^2 = 29.75 + (-1.5)^2 + (1)^2 + \cdots = 88.25 \\
L'(\phi_B) &= 65.0 \\
L'(\phi_C) &= 257.0
\end{aligned}
$$

   The best set is ϕ_B.

3. Discuss the procedure you would implement on a computer to find the optimal set of parameters, had the table above not been provided.

   If the table is not provided, it is intractable to evaluate and compare all possible sets of parameters. However, we still want to minimize the MLE (or MAP) objective $L(\phi)$. We can improve a set of parameters ϕ by following the opposite of the objective's gradient $\nabla_\phi L(\phi)$, *i.e.* performing *gradient descent* steps

$$\phi \leftarrow \phi - \gamma \nabla_\phi L(\phi),$$

where γ is the learning rate. As alternatives to gradient descent, one could mention genetic or Markov chain Monte Carlo (MCMC) algorithms.

4. Using the heuristic $h(x, y; \phi_D)$ with $\phi_D = (-2, 1, 5, 0.5, -2, 7)$, apply 5 iterations of the greedy search algorithm, starting from room $(1, 1)$.

   The greedy search algorithm uses a priority queue as fringe. Our goal is to reach the exit as fast as possible. As our heuristic approximates the time to reach the exit from room (x, y), we should prioritize rooms that have the lowest heuristic in our queue. We keep in our fringe the rooms that are reachable from the visited (closed set) rooms. For convenience, we denote rooms as a two-digit number $10 \times x + y$, annotate rooms with their priority $h(x, y; \phi_D)$ and prune visited rooms from the fringe.

Fringe (priority queue)	Closed
11(9.5)	
21(8.5) 12(8.0)	11
21(8.5) 22(7.0) 13(6.5)	11 12
21(8.5) 14(7.0) 22(7.0) 23(5.5)	11 12 13
21(8.5) 14(7.0) 22(7.0) 24(5.0) 33(4.5)	11 12 13 23
21(8.5) 14(7.0) 22(7.0) 32(6.0) 24(5.0) 43(3.5) 34(3.0)	11 12 13 23 33

Supplementary materials

- Heteroscedasticity
- Laplace distribution
- Chapter 18 of the reference textbook.