Machine Learning

Bayesian methods

Wojciech Czech

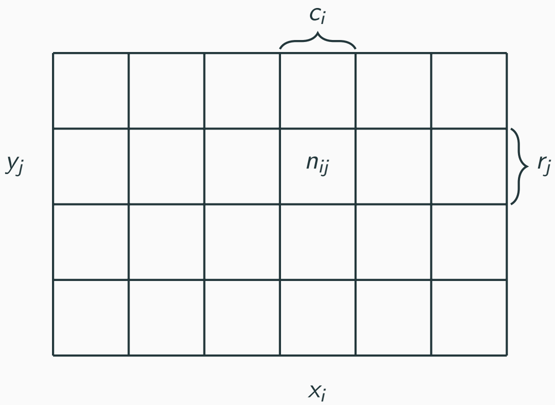
February 1, 2025

Institute of Computer Science

Table of contents

The Rules of Probability

Sample space of two random variables



Random variable X can take any of the values x_i, where $i = 1, \dots, M$
Random variable Y can take any of the values y_j, where $j = 1, \dots, L$

Symbols

- N - number of trials, in which we sample both X and Y
- n_{ij} - number of trials, in which we observed $X = x_i$ and $Y = y_j$
- c_i - number of trials, in which we observed $X = x_i$ regardless of value taken by Y
- r_j - number of trials, in which we observed $Y = y_j$ regardless of value taken by X
- We consider limit $N \to \infty$

Sum rule

Joint probability: $p(X = x_i, Y = y_j) = \dfrac{n_{ij}}{N}$

Sum rule

Joint probability: $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$

Marginal probability: $p(X = x_i) = \frac{c_i}{N}$

Sum rule

Joint probability: $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$

Marginal probability: $p(X = x_i) = \frac{c_i}{N}$

$$c_i = \sum_j n_{ij}$$

Sum rule

Joint probability: $p(X = x_i, Y = y_j) = \dfrac{n_{ij}}{N}$

Marginal probability: $p(X = x_i) = \dfrac{c_i}{N}$

$$c_i = \sum_j n_{ij}$$

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_j n_{ij} = \sum_j \frac{n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j)$$

Sum rule

Joint probability: $p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$

Marginal probability: $p(X = x_i) = \frac{c_i}{N}$

$$c_i = \sum_j n_{ij}$$

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_j n_{ij} = \sum_j \frac{n_{ij}}{N} = \sum_j p(X = x_i, Y = y_j)$$

Sum rule of probability: $p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$

Product rule

Conditional probability: $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$

Product rule

Conditional probability: $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$

Only trials for which $X = x_i$

Product rule

Conditional probability: $p(Y = y_j | X = x_i) = \dfrac{n_{ij}}{c_i}$

Only trials for which $X = x_i$

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p(Y = y_j | X = x_i)p(X = x_i)$$

Product rule

Conditional probability: $p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$

Only trials for which $X = x_i$

$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N} = p(Y = y_j | X = x_i)p(X = x_i)$

Product rule of probability:

$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i)p(X = x_i)$

Rules of probability

Sum rule

$$p(X) = \sum_{Y} p(X, Y)$$

Product rule

$$p(X, Y) = p(Y|X)p(X)$$

Symmetry property

$$p(X, Y) = p(Y, X)$$

Example 1

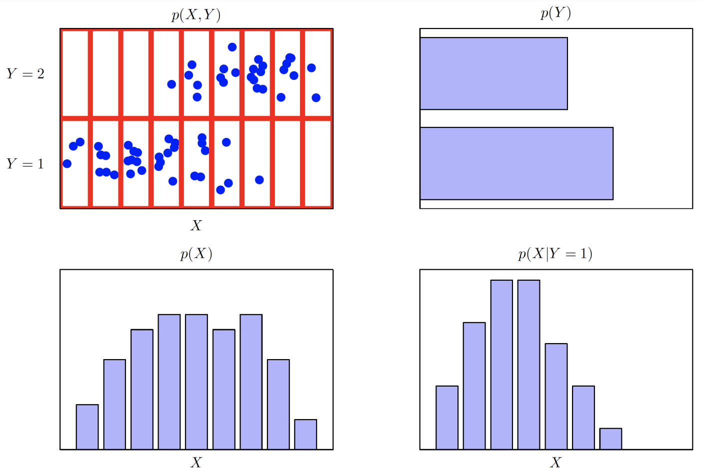


Figure 1: Distribution over two random variables [1].

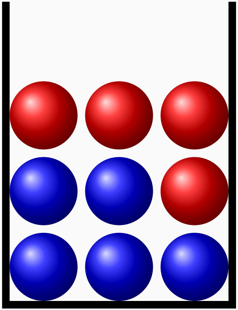
Bayes' theorem

Bayes' theorem

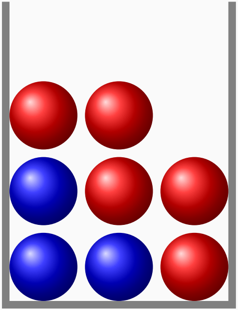
$$p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(Y,X)}{p(X)} = \frac{p(X|Y)p(Y)}{p(X)} \tag{1}$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(Y,X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)} \tag{2}$$

Example 2: Colored balls



$p(B = black) = 6/10$



$p(B = gray) = 4/10$

Example 2: Probabilites

$$p(C = blue|B = black) = 5/9$$

$$p(C = red|B = black) = 4/9$$

$$p(C = blue|B = gray) = 3/8$$

$$p(C = red|B = gray) = 5/8$$

$$p(C = red) = p(C = red, B = black) + p(C = red, B = gray) =$$

$$= p(C = red|B = black)p(B = black) + p(C = red|B = gray)p(B = gray)$$

$$p(C = red) = 31/60$$

$$p(C = blue) = 1 - p(C = red) = 29/60$$

Example 2: Probabilites

$$p(B = black|C = red) = \frac{p(C = red|B = black)p(B = black)}{p(C = red)} = 48/93$$

prior: 0.6, posterior: 0.52

$$p(B = gray|C = red) = 1 - p(B = black|C = red) = 45/93$$

$$p(B = black|C = blue) = \frac{p(C = blue|B = black)p(B = black)}{p(C = blue)} = 60/87$$

prior: 0.6, posterior: 0.69

$$p(B = gray|C = blue) = 1 - p(B = black|C = blue) = 27/87$$

Prior and posterior probability

Prior probability

Probability available before observation, e.g., $p(B)$

Posterior probability

Probability obtained after observation, e.g., $p(B|C)$

Bayes' theorem converts a prior probability into a posterior probability by incorporating the evidence provided by the observed data.

Bayesian probability

Classical probability (Frequentist probability)

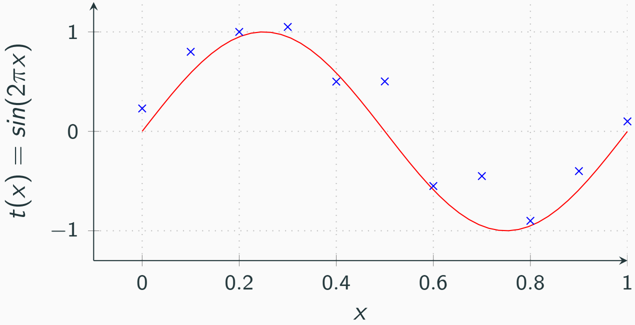
Limit of an event relative frequency in many trials

Bayesian probability

Quantifies uncertainity in the light of new evidence

Bayesian curve fitting

Example 3: Curve fitting

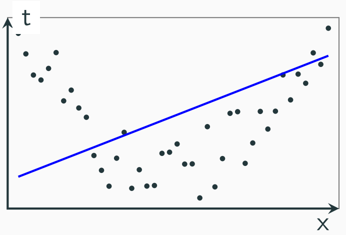


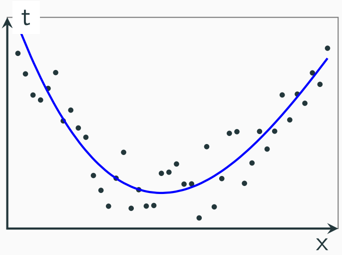
$$\boldsymbol{x} = [x_1, \ldots, x_n]^T, \boldsymbol{t} = [t_1, \ldots, t_n]^T \tag{3}$$

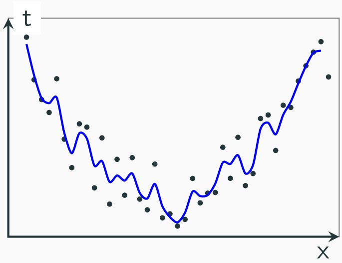
$$y(x, \boldsymbol{w}) = \sum_{i=0}^{m} w_i x^i \tag{4}$$

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{i=0}^{m} (y(x_i, \boldsymbol{w}) - t_i)^2 \tag{5}$$

Model complexity vs. size of training set







Example 3: Curve fitting - regularization

Model complexity and overfitting

- Number of model parameters vs. size of training set
- Values of model parameters

Regularization

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=0}^{m}(y(x_i, \boldsymbol{w}) - t_i)^2 + \frac{\lambda}{2}\left\|\boldsymbol{w}\right\|^2 \tag{6}$$

Example 3: Curve fitting - Bayesian formulation

- Assumptions regarding $\boldsymbol{w}$ reflected by $p(\boldsymbol{w})$ - prior probability distribution
- $\mathcal{D} = \{t_1, \ldots, t_n\}$ - observed data
- Measure uncertainty in $\boldsymbol{w}$ after observing $\mathcal{D}$ using posterior probability $p(\boldsymbol{w}|\mathcal{D})$
- $p(\mathcal{D}|\boldsymbol{w})$ - likelihood function[1] (function of parameter vector $\boldsymbol{w}$)

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})} \tag{7}$$

- Bayes' theorem

$$\textit{posterior} \propto \textit{likelihood} \times \textit{prior}$$

[1] How probable the observed data set is for different settings of $\boldsymbol{w}$

Likelihood function

Likelihood function

Measures fit of a statistical model to observed data for a given values of model parameters. Interpreted as a function of parameters only, treating random variables as fixed. Formed from joint probability distribution of the sample.

Maximum likelihood estimation

Finding combination of model parameter values that maximize the probability of drawing the sample obtained.

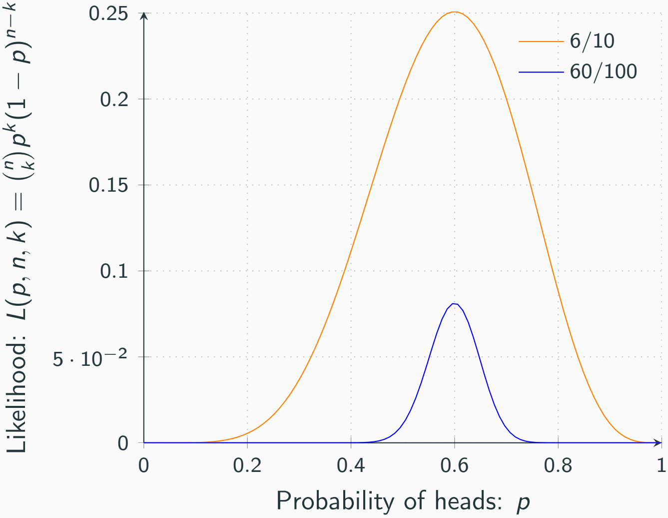
Log likelihood function

Maximizing the likelihood is equivalent to maximizing the log-likelihood and minimizing negative log-likelihood function.

Example 4: Coin flip

- Flipping coin 10 times, obtained 6 heads and 4 tails
- Binomial distribution: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- H_1: Fair coin $p(heads) = w_1 = 0.5$
- $P(X = 6|w_1) \approx 0.21$
- H_2: Trick coin $p(heads) = w_2 = 0.75$
- $P(X = 6|w_2) \approx 0.15$
- $L(H_1) = 0.21$, $L(H_2) = 0.15$
- Likelihood ratio: $LR = \frac{L(H_1)}{L(H_2)} = 1.4$
- Hypothesis H_1 explains the data better

Example 4: Coin flip - maximum likelihood



Example 5: Students

Piotr i Paweł spierają się, czy studenci z ich roku preferują Bayesowską (H_1) czy częstotliwościową (H_2) interpretację prawdopodobieństwa. Aby rozstrzygnąć spór, poprosili koleżankę Kasię żeby zapytała 10 losowych studentów opuszczających salę po zajęciach o ich zdanie w spornej kwestii. Dwaj oponenci zgodzili się również wstępnie, że można uznać, że odsetek studentów preferujących interpretację Bayesowską θ podlega rozkładowi $Beta(\theta; \alpha, \beta) = c\, \theta^{\alpha-1}(1-\theta)^{\beta-1}$, przy czym $\alpha = 2$ i $\beta = 2$, a c jest stałą normalizującą, którą można obliczyć za pomocą wzoru: $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, przy czym $\Gamma(m) = (m-1)!$ jeśli m jest liczbą naturalną.

a) Niech k oznacza liczbę studentów w próbce zbadanej przez Kasię, preferujących Bayesowską interpretację prawdopodobieństwa. Jaka jest funkcja masy prawdopodobieństwa dla **dyskretnej** zmiennej k?

b) Oblicz rozkład *a posteriori* dla θ jeśli $k = 6$.

c) Znajdź rozwiązanie MAP (maximum a posteriori) dla parametru θ. Czyja hipoteza (H_1 vs. H_2) jest bardziej potwierdzona przez obserwowane dane?

Example 5: solution (1)

- H_1: $\theta > 0.5$ (Piotr)
- H_2: $\theta \leqslant 0.5$ (Paweł)

a) $p(k) = \binom{10}{k}\theta^k(1-\theta)^{10-k}$ (Kasia odpytuje wychodzących)

b) $p(\theta) = c\,\theta(1-\theta)$ (**prior**, zgodzili się wstępnie)
$p(k=6|\theta) = \binom{10}{6}\theta^6(1-\theta)^4$ (**likelihood**)
$p(\theta|k=6) \propto \binom{10}{6}\theta^6(1-\theta)^4 c\,\theta(1-\theta) \propto \theta^7(1-\theta)^5$ (**posterior**, nieznormalizowana)

Można zauważyć, że rozkład posterior jest rozkładem $Beta(\theta; 8, 6)$ czyli, że: $\alpha = 8$, $\beta = 6$.

Zamiast całkować, skorzystamy ze wzoru: $c = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{13!}{7!5!} = 10296$
$p(\theta|k=6) = \frac{13!}{7!5!}\theta^7(1-\theta)^5$

Example 5: solution (2)

c)

$$\frac{\partial}{\partial\theta}\theta^7(1-\theta)^5 = 0$$

$$\theta^6(1-\theta)^4(7-12\theta) = 0$$

$$\theta = 0 \text{ lub } \theta = \tfrac{7}{12}$$

Maximum dla $\theta = \frac{7}{12} \approx 0.58$

Dane bardziej wspierają hipotezę Piotra

Bayesian curve fitting

$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1}) \tag{8}$$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{9}$$

Precision: $\beta^{-1} = \sigma^2$

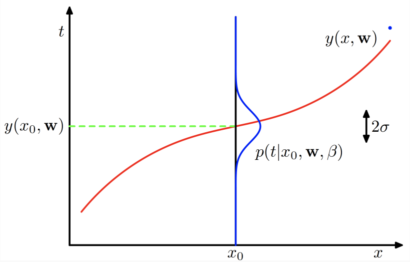


Figure 2: See: [1].

Bayesian curve fitting - model parameters

$$\boldsymbol{x} = [x_1, \dots, x_n]^T, \boldsymbol{t} = [t_1, \dots, t_n]^T \tag{10}$$

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = \prod_{i=0}^{n} \mathcal{N}(t_i | y(x_i, \boldsymbol{w}), \beta^{-1}) \tag{11}$$

$$\ln p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{i=0}^{n} (y(x_i, \boldsymbol{w}) - t_i)^2 + \frac{n}{2} \ln \beta - \frac{n}{2} \ln (2\pi) \tag{12}$$

- Data drawn independently from distribution Eq. 8 (i.i.d.)
- Maximize likelihood with respect to $\boldsymbol{w}$ - maximize log likelihood - minimize negative log likelihood - mimimize root square error function (see: Equation 5).
- Sum-of-squares error function appears as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution.

Bayesian curve fitting - maximum likelihood

Maximum likelihood parameters

$$\boldsymbol{w}_{ML} = \operatorname*{argmin}_{\boldsymbol{w}} \sum_{i=0}^{n} (y(x_i, \boldsymbol{w}) - t_i)^2 \tag{13}$$

$$\frac{1}{\beta_{ML}} = \sum_{i=1}^{n} (y(x_i, \boldsymbol{w}_{ML}) - t_i)^2 \tag{14}$$

Predictive distribution

$$p(t|x, \boldsymbol{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, \boldsymbol{w}_{ML}), \beta_{ML}^{-1}) \tag{15}$$

Bayesian curve fitting - maximum posterior

Prior distribution over polynomial coefficients

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) = \left(\frac{\alpha}{2\pi}\right)^{(m+1)/2} \exp\left\{-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\} \tag{16}$$

Bayes' theorem

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha) \tag{17}$$

MAP - maximium posterior
Determine most probable value of $\boldsymbol{w}$ given data

$$\boldsymbol{w}_{MP} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{\frac{\beta}{2}\sum_{i=0}^{n}(y(x_i, \boldsymbol{w}) - t_i)^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\} \tag{18}$$

See Equation 6 and Equation 18 : $\lambda = \alpha/\beta$

Posterior predictive distribution

$$p(t|x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t|x, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}) d\boldsymbol{w} \tag{19}$$

$p(t|x, \boldsymbol{w})$ - see Equation 8

$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t})$ - posterior distribution over parameters, see Equation 17

$$p(t|x, \boldsymbol{x}, \boldsymbol{t}) = \mathcal{N}(t|m(x), s^2(x)) \tag{20}$$

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S} \sum_{i=0}^{n} \boldsymbol{\phi}(x_i) t_i \tag{21}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^T \boldsymbol{S} \boldsymbol{\phi}(x) \tag{22}$$

$$\boldsymbol{S}^{-1} = \alpha \boldsymbol{I} + \beta \sum_{i=0}^{n} \boldsymbol{\phi}(x_i) \boldsymbol{\phi}(x_i)^T \tag{23}$$

$$\boldsymbol{\phi}(x) = [1, x, \ldots, x^m]^T \tag{24}$$

Bayesian inference

Example 6: Binary classification problem

- $\boldsymbol{x}$ - input feature vector
- t - output target variable: $t \in \{\mathcal{C}_1, \mathcal{C}_2\}$, alternatively $t \in [0, 1]$
- Inference problem - determining $p(\boldsymbol{x}, t)$
- Inference problem vs. decision problem
- Intuition: select the class having the higher posterior probability

$$p(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\boldsymbol{x})} \tag{25}$$

Minimizing misclassification rate

Decision region $\mathcal{R}_k$

All points $\boldsymbol{x}$ belonging to $\mathcal{R}_k$ are assinged to class $\mathcal{C}_k$

Decision boundaries

The boundaries between decision regions

$$p(mistake) = p(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1) \tag{26}$$

$$p(mistake) = \int_{\mathcal{R}_1} p(\boldsymbol{x}, \mathcal{C}_2) d\boldsymbol{x} + \int_{\mathcal{R}_2} p(\boldsymbol{x}, \mathcal{C}_1) d\boldsymbol{x} \tag{27}$$

Decision rule

If $p(\boldsymbol{x}, \mathcal{C}_1) > p(\boldsymbol{x}, \mathcal{C}_2)$, then assign $\boldsymbol{x}$ to class $\mathcal{C}_1$
If $p(\mathcal{C}_1|\boldsymbol{x}) > p(\mathcal{C}_2|\boldsymbol{x})$, then assign $\boldsymbol{x}$ to class $\mathcal{C}_1$
Select largest posterior probability to minimize probability of making a mistake

Decision boundary

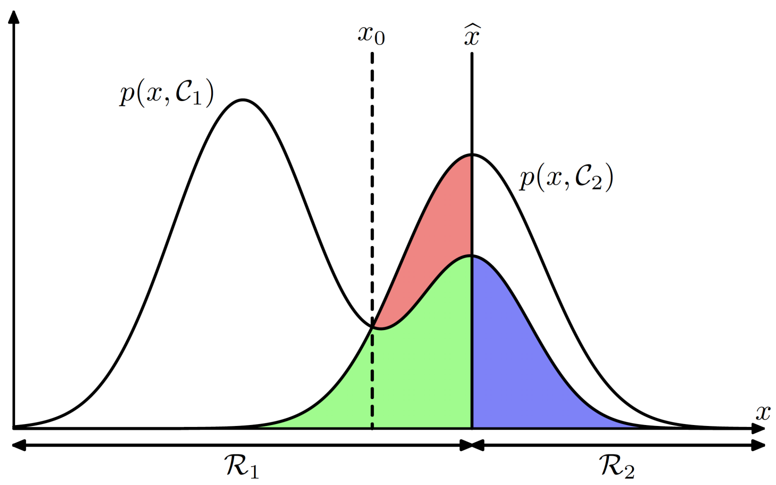


Figure 3: Decision boundary [1].

Minimizing the expected loss

Loss (cost) function

Overall measure of loss incurred in taking any of the available decisions or actions

Loss marix $\boldsymbol{L}$

$\boldsymbol{L}_{kj}$ - the cost of assigning $\boldsymbol{x}$ to class j, while the true class is k

$$\mathbb{E}[\boldsymbol{L}] = \sum_k \sum_j \int_{\mathcal{R}_j} \boldsymbol{L}_{kj} p(\boldsymbol{x}, \mathcal{C}_k) d\boldsymbol{x} \tag{28}$$

Decision rule

Assign new $\boldsymbol{x}$ to class j for which the quantity $\sum_k L_{kj} p(\mathcal{C}_k|\boldsymbol{x})$ is a minimum

Classification problem: inference and decision

Inference stage

Use training data to learn a model for $p(\mathcal{C}_k|\boldsymbol{x})$

Decision stage

Use posterior probabilities to assign classes

Discriminat function

Solve inference and decision together by learning discriminant function, which maps $\boldsymbol{x}$ directly to decisions

Generative models

- Determine class conditional probability densities $p(\boldsymbol{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$
- Determine class prior probabilities $p(\mathcal{C}_k)$
- Use Bayes' theorem to find posterior probabilities $p(\mathcal{C}_k|\boldsymbol{x})$
- Use decision theory to determine class membership for each new input $\boldsymbol{x}$
- Sampling model can generate new synthetic data

Discriminative models

- Determine posterior probabilities $p(\mathcal{C}_k|\boldsymbol{x})$ directly
- Use decision theory to determine class membership for each new input $\boldsymbol{x}$

Discriminant function models

- Find function $f(\boldsymbol{x})$, which maps input vector $\boldsymbol{x}$ directly to class label
- $f : \boldsymbol{x} \rightarrow \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

Advantages of knowing posterior probabilities[2]

Minimizing risk

Easy revision of loss matrix

Reject option

Exclude regions with big uncertainty about class membership. Simpler adjustment of reject criteria.

Compensating for class priors

Adjusting to imbalanced training datasets

Combining models

Utilizing conditional independence property

[2] vs. knowing discrimitative function only

Naive Bayes Classifier

- Divide classification problem into subproblems with different input spaces
- Obtain posterior probabilities for the classes for each input space separately
- Combine output using rules of probability (normalization)
- **Assume that inputs are independent (conditional independence)**

$$p(\boldsymbol{x}_I, \boldsymbol{x}_B | \mathcal{C}_k) = p(\boldsymbol{x}_I | \mathcal{C}_k) p(\boldsymbol{x}_B | \mathcal{C}_k) \tag{29}$$

$$\begin{aligned} p(\mathcal{C}_k | \boldsymbol{x}_I, \boldsymbol{x}_B) \propto p(\boldsymbol{x}_I, \boldsymbol{x}_B | \mathcal{C}_k) p(\mathcal{C}_k) &\propto p(\boldsymbol{x}_I | \mathcal{C}_k) p(\boldsymbol{x}_B | \mathcal{C}_k) p(\mathcal{C}_k) \\ &\propto \frac{p(\mathcal{C}_k | \boldsymbol{x}_I) p(\mathcal{C}_k | \boldsymbol{x}_B)}{p(\mathcal{C}_k)} \end{aligned} \tag{30}$$

Naive Bayes Classifier - multiple features

$$p(\mathcal{C}_k|x_1,\dots,x_n) = \frac{p(x_1|\mathcal{C}_k)p(x_2|\mathcal{C}_k)\dots p(x_n|\mathcal{C}_k)p(\mathcal{C}_k)}{p(x_1)p(x_2)\dots p(x_n)} \tag{31}$$

$$p(\mathcal{C}_k|x_1,\dots,x_n) = \frac{p(\mathcal{C}_k)\prod_{i=1}^{n} p(x_i|\mathcal{C}_k)}{\prod_{i=1}^{n} p(x_i)} \tag{32}$$

$$p(\mathcal{C}_k|x_1,\dots,x_n) \propto p(\mathcal{C}_k)\prod_{i=1}^{n} p(x_i|\mathcal{C}_k) \tag{33}$$

$$\mathcal{C} = \operatorname*{argmax}_{\mathcal{C}_k}\left\{p(\mathcal{C}_k)\prod_{i=1}^{n} p(x_i|\mathcal{C}_k)\right\} \tag{34}$$

Example 7: Weather conditions for playing golf

	Outlook x_1	Temperature x_2	Humidity x_3	Windy x_4	Play Golf
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

Example 7: Prior

	No. cases	$p(\mathcal{C}_k)$
Yes ($\mathcal{C}_1$)	9	9/14
No ($\mathcal{C}_2$)	5	5/14
Total	14	1

Example 7: Likelihood

	Yes ($\mathcal{C}_1$)	No ($\mathcal{C}_2$)	$p(x_1\|\mathcal{C}_1)$	$p(x_1\|\mathcal{C}_2)$
Sunny	3	2	1/3	2/5
Overcast	4	0	4/9	0
Rainy	2	3	2/9	3/5
Total	9	5	1	1

	Yes ($\mathcal{C}_1$)	No ($\mathcal{C}_2$)	$p(x_2\|\mathcal{C}_1)$	$p(x_2\|\mathcal{C}_2)$
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	1/3	1/5
Total	9	5	1	1

	Yes ($\mathcal{C}_1$)	No ($\mathcal{C}_2$)	$p(x_3\|\mathcal{C}_1)$	$p(x_3\|\mathcal{C}_2)$
High	3	4	1/3	4/5
Normal	6	1	2/3	1/5
Total	9	5	1	1

	Yes ($\mathcal{C}_1$)	No ($\mathcal{C}_2$)	$p(x_4\|\mathcal{C}_1)$	$p(x_4\|\mathcal{C}_2)$
False	6	2	2/3	2/5
True	3	3	1/3	3/5
Total	9	5	1	1

Example 7: Posterior

Query: $\boldsymbol{x}_q = [x_1, x_2, x_3, x_4] = [\mathit{Sunny}, \mathit{Hot}, \mathit{Normal}, \mathit{False}]$

$p(\mathcal{C}_1|\boldsymbol{x}_q) \propto \frac{1}{3}\frac{2}{9}\frac{2}{3}\frac{2}{3}\frac{9}{14} = 0.02116$

$p(\mathcal{C}_2|\boldsymbol{x}_q) \propto \frac{2}{5}\frac{2}{5}\frac{1}{5}\frac{2}{5}\frac{5}{14} = 0.00457$

$p(\mathcal{C}_1|\boldsymbol{x}_q) = \frac{0.02116}{0.02116+0.00457} = 0.822$

$p(\mathcal{C}_2|\boldsymbol{x}_q) = 1 - 0.822 = 0.178$

Prediction $\mathcal{C}_1$: Golf will be played

Gaussian Naive Bayes classifier

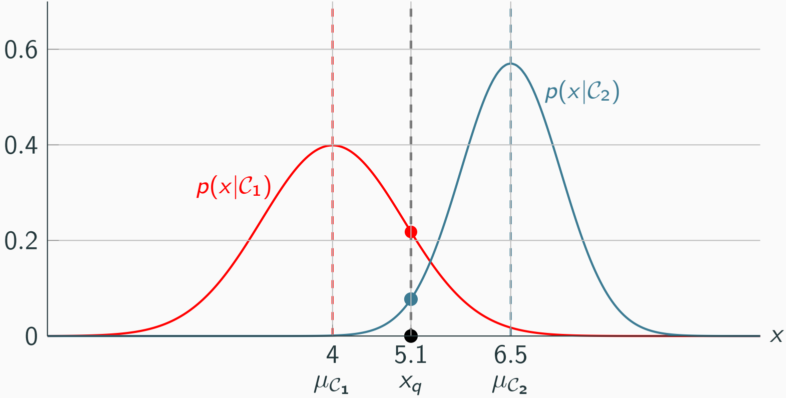
- x_i has continuous values
- Assume that $p(x_i|\mathcal{C}_k)$ is Gaussian

$$p(x_i|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi\sigma^2_{\mathcal{C}_k}}} \exp\left\{-\frac{(x_i - \mu_{\mathcal{C}_k})^2}{2\sigma^2_{\mathcal{C}_k}}\right\} \tag{35}$$

- $\mu_{\mathcal{C}_k}$, $\sigma^2_{\mathcal{C}_k}$ - mean and variance of continuous x_i calculated for a given class $\mathcal{C}_k$

See: Box-Cox or Yeo-Johnson transformations

Gaussian Naive Bayes classifier - visualization



Example 8: Classify gender of the person

	Height x_1	Weight x_2	Foot size x_3	Gender
1	6	180	12	Male
2	5.92	190	11	Male
3	5.58	170	12	Male
4	5.92	165	10	Male
5	5	100	6	Female
6	5.5	150	8	Female
7	5.42	130	7	Female
8	5.75	150	9	Female

Example 8: Means and variances

Gender	μ_{x_1}	$\sigma^2_{x_1}$	μ_{x_2}	$\sigma^2_{x_2}$	μ_{x_3}	$\sigma^2_{x_3}$
Male	5.885	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
Female	5.415	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

$$Male : \mathcal{C}_1,\ Female : \mathcal{C}_2$$
$$p(\mathcal{C}_1) = 0.5,\ p(\mathcal{C}_2) = 0.5$$
$$\text{Query: } \boldsymbol{x}_q = [x_1, x_2, x_3] = [6, 130, 8]$$

Example 8: Likelihood $\mathcal{C}_1$

$$p(\mathcal{C}_1|\boldsymbol{x}_q) \propto p(\mathcal{C}_1)\prod_{i=1}^{3} p(x_i|\mathcal{C}_1)$$

$$p(x_1 = 6|\mathcal{C}_1) = \frac{1}{\sqrt{2\pi\sigma_{x_1}^2}} \exp\left\{-\frac{(6-\mu_{x_1})^2}{2\sigma_{x_1}^2}\right\} \approx 1.57889$$

$$p(x_2 = 130|\mathcal{C}_1) = \frac{1}{\sqrt{2\pi\sigma_{x_2}^2}} \exp\left\{-\frac{(130-\mu_{x_2})^2}{2\sigma_{x_2}^2}\right\} \approx 5.9881 \times 10^{-6}$$

$$p(x_3 = 8|\mathcal{C}_1) = \frac{1}{\sqrt{2\pi\sigma_{x_3}^2}} \exp\left\{-\frac{(8-\mu_{x_3})^2}{2\sigma_{x_3}^2}\right\} \approx 1.3112 \times 10^{-3}$$

$$p(\mathcal{C}_1|\boldsymbol{x}_q) \propto 6.1984 \times 10^{-9}$$

Example 8: Likelihood $\mathcal{C}_2$

$$p(\mathcal{C}_2|\boldsymbol{x}_q) \propto p(\mathcal{C}_2)\prod_{i=1}^{3} p(x_i|\mathcal{C}_2)$$

$$p(x_1 = 6|\mathcal{C}_2) \approx 2.2346 \times 10^{-1}$$

$$p(x_2 = 130|\mathcal{C}_2) \approx 1.6789 \times 10^{-2}$$

$$p(x_3 = 8|\mathcal{C}_2) \approx 2.8669 \times 10^{-1}$$

$$p(\mathcal{C}_2|\boldsymbol{x}_q) \propto 5.3778 \times 10^{-4}$$

Example 8: Posterior

$$p(\mathcal{C}_1|\boldsymbol{x}_q) = \frac{6.1984 \times 10^{-9}}{6.1984 \times 10^{-9} + 5.3778 \times 10^{-4}} = 1.1526 \times 10^{-5}$$

$$p(\mathcal{C}_2|\boldsymbol{x}_q) = 1 - 1.1526 \times 10^{-5} \approx 1$$

Prediction $\mathcal{C}_2$: Female

Naive Bayes classifier - advantages

- If the assumption about conditional independence of inputs is true - performs better than other models
- Requires relatively small sizes of training data, fast training
- Easy to implement
- Each distribution can be independently estimated as a one-dimensional distribution (avoiding problems with curse of dimensionality)
- Suitable for multiclass classification
- Simple update of probabilities after new data is received
- Ability to use more complex distributions for continuous input

Naive Bayes classifier - disadvantages

- Very strong assumption of conditional independence of inputs
- Zero frequency problem
- Weights equally all input features
- Problems with handling countinuous variables (discretization vs. assumptions regarding probability distribution)
- Problems with handling imbalanced datasets
- Bad estimator of probability (independence assumption is false if you are trying to estimate the probability)

Naive Bayes classifier - hints

- Reduce highly correlated features (e.g. using PCA)
- Execute domain specific feature engineering
- Try introducing more realistic prior probabilities (domain knowledge)
- Use ensemble methods: bagging or boosting to reduce variance (improve generalization capabilities)
- Use Box-Cox or Yeo-Johnson transformations to make features more Gaussian
- Use **Laplace smoothing** to handle records with zero values in x_i dimension

Naive Bayes classifier - applications

- Real time prediction
- Multi-class predictions
- Text classification
- Spam filtering
- Sentiment analysis
- Recommendation systems

Multivariate Gaussian distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{36}$$

$\boldsymbol{x} \in \mathbb{R}^{D\times 1}$

$\boldsymbol{\mu} \in \mathbb{R}^{D\times 1}$ - mean vector

$\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ - covariance matrix

$|\boldsymbol{\Sigma}|$ - determinant of $\boldsymbol{\Sigma}$

$\boldsymbol{\Delta}^2 = (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ - Mahalanobis distance

Multivariate Gaussian distribution - properties

- Constant on surfaces in $\boldsymbol{x}$-space for which Mahalanobis distance is constant
- Elliptical surface of constant probability density for a 2D Gaussian
- Eigendecomposition of covariance matrix

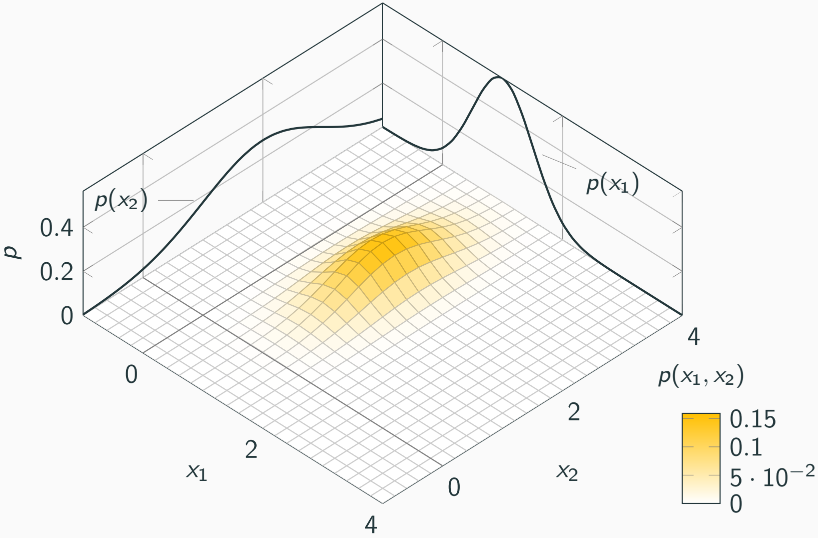
$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{37}$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{u}_i \boldsymbol{u}_i^T \tag{38}$$

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{i=1}^{D} \lambda_i^{1/2} \tag{39}$$

Multivariate Gaussian distribution - properties

- Number of free parameters $D(D+3)/2$
- Covariance matrix restricted to diagonal: $2D$ parameters
- Covariance matrix restricted to scaled identity $D+1$ parameters
- Unimodal distribution (single maximum)



$p(x_2)$
$p(x_1)$
0.4
0.2
0
p
0
2
4
x_1
0
2
4
x_2
$p(x_1, x_2)$
0.15
0.1
$5 \cdot 10^{-2}$
0

Partitioned Gaussian distributions

Two sets of variables

If two sets of variables are jointly Gaussian, then conditional distribution of one set conditioned on the other is also Gaussian. Marginal distribution of either set is also Gaussian.

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix} \tag{40}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T, \boldsymbol{\Sigma}_{ab} = \boldsymbol{\Sigma}_{ba}^T \tag{41}$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}, \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} \tag{42}$$

Partitioned Gaussians

Conditional distribution

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \tag{43}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b) \tag{44}$$

Marginal distribution

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \tag{45}$$

Bayesian linear regression

Linear basis function model

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{m-1} w_i \phi_i(\boldsymbol{x}) \tag{46}$$

$\phi_i(\boldsymbol{x})$ - basis function, w_0 - bias

$\phi_0(\boldsymbol{x}) = 1$

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=0}^{m-1} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\mathrm{x}) \tag{47}$$

$$\boldsymbol{w} = [w_0, \ldots, w_{m-1}]^T$$

$$\boldsymbol{\phi} = [\phi_0, \ldots, \phi_{m-1}]^T$$

Non-linear basis functions

- Polynomial: $\phi_i(x) = x^i$
- Gaussian: $\phi_i(x) = \exp\left\{-\frac{(x-\mu_i)^2}{2s^2}\right\}$
- Sigmoidal: $\phi_i(x) = \sigma\left(\frac{x-\mu_i}{s}\right)$, $\sigma(a) = \frac{1}{1+\exp(-a)}$
- Fourier
- Wavelet

Bayesian linear basis function model - formulation

$$t(\boldsymbol{x}) = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon \tag{48}$$

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) = \mathcal{N}(t|\boldsymbol{w}^T\phi(\boldsymbol{x}), \beta^{-1}) \tag{49}$$

Bayesian linear basis function model - likelihood

$\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ - set of multidimensional inputs

$$\boldsymbol{t} = [t_1, \ldots, t_n]^T \tag{50}$$

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T \phi(\boldsymbol{x}), \beta^{-1}) \tag{51}$$

$$\ln p(\boldsymbol{t}|\boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{i=0}^{n} (t_i - \boldsymbol{w}^T \phi(\boldsymbol{x}))^2 + \frac{n}{2} \ln \beta - \frac{n}{2} \ln (2\pi) \tag{52}$$

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{i=0}^{n} (t_i - \boldsymbol{w}^T \phi(\boldsymbol{x}))^2 \tag{53}$$

Bayesian linear basis function model - maximum likelihood

Maximum likelihood parameters

$$\nabla \ln p(\boldsymbol{t}|\boldsymbol{w}, \beta) = 0 \tag{54}$$

$$\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{t} \tag{55}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{w}_{ML}^T \phi(\boldsymbol{x}_i) - t_i)^2 \tag{56}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_{m-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \dots & \phi_{m-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_n) & \phi_1(\boldsymbol{x}_n) & \dots & \phi_{m-1}(\boldsymbol{x}_n) \end{bmatrix} \tag{57}$$

Moore-Penrose pseudo-inverse

Bayesian linear basis function model - maximum posterior

Prior distribution over model parameters

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0) \tag{58}$$

Posterior distribution

$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_n, \boldsymbol{S}_n) \tag{59}$$

$$\boldsymbol{m}_n = \boldsymbol{S}_n(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\Phi^T\boldsymbol{t}) \tag{60}$$

$$\boldsymbol{S}_n^{-1} = \boldsymbol{S}_0^{-1} + \beta\Phi^T\Phi \tag{61}$$

MAP - maximium posterior
Determine most probable value of $\boldsymbol{w}$ given data (mode equals mean)

$$\boldsymbol{w}_{MP} = \boldsymbol{m}_n \tag{62}$$

Zero-mean isotropic Gaussian prior

Prior distribution

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \tag{63}$$

Posterior distribution

$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_n, \boldsymbol{S}_n) \tag{64}$$

$$\boldsymbol{m}_n = \beta \boldsymbol{S}_n \Phi^T \boldsymbol{t} \tag{65}$$

$$\boldsymbol{S}_n^{-1} = \alpha \boldsymbol{I} + \beta \Phi^T \Phi \tag{66}$$

MAP - maximium posterior

$$\ln p(\boldsymbol{w}|\boldsymbol{t}) = -\frac{\beta}{2}\sum_{i=1}^{n}(t_i - \boldsymbol{w}^T\phi(\boldsymbol{x}_i))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + const \tag{67}$$

Online Bayesian regression

- Sequence of :
  - Feature vectors: $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_i, \dots$
  - Target variables $t_1, t_2, \dots, t_i, \dots$
  - Model parameters: $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_i, \dots$

$$\boldsymbol{w}_i \xrightarrow{t_i} \boldsymbol{w}_{i+1}$$

- Current prior distribution of model parameters $p(\boldsymbol{w}_i)$
- **Predictive distribution** $p(t_i|\boldsymbol{x}_i)$
- Likelihood $p(t_i|\boldsymbol{x}_i, \boldsymbol{w}_i)$
- Posterior distribution $p(\boldsymbol{w}_{i+1}|\boldsymbol{w}_i, \boldsymbol{x}_i, t_i)$

$$p(\boldsymbol{w}_{i+1}|\boldsymbol{w}_i, \boldsymbol{x}_i, t_i) \propto p(t_i|\boldsymbol{x}_i, \boldsymbol{w}_i)p(\boldsymbol{w}_i) \tag{68}$$

Online Bayesian regression - predictive distribution

$$p(t_i|\boldsymbol{x}_i) = \int p(t_i|\boldsymbol{x}_i, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} \tag{69}$$

$$p(t_i|\boldsymbol{x}_i) \propto p(t_i|\boldsymbol{x}_i, \boldsymbol{w}_i)p(\boldsymbol{w}_i) \tag{70}$$

Online belief updating

From Eq. 70 and Eq. 68, assuming initial prior $p(\boldsymbol{w}_0)$:

$$p(t_0|\boldsymbol{x}_0) \propto p(t_0|\boldsymbol{x}_0, \boldsymbol{w}_0)p(\boldsymbol{w}_0) \tag{71}$$

After first observation:

$$p(\boldsymbol{w}_1|\boldsymbol{w}_0, \boldsymbol{x}_0, t_0) \propto p(t_0|\boldsymbol{x}_0, \boldsymbol{w}_0)p(\boldsymbol{w}_0) \tag{72}$$

Predictive distribution:

$$p(t_1|\boldsymbol{x}_1) \propto p(t_1|\boldsymbol{x}_1, \boldsymbol{w}_1)\underbrace{p(t_0|\boldsymbol{x}_0, \boldsymbol{w}_0)p(\boldsymbol{w}_0)}_{p(\boldsymbol{w_1})} \tag{73}$$

The prior of the weights for the current iteration is the posterior of the weights at the previous iteration.

Online belief updating

After second observation:

$$p(\boldsymbol{w}_2|\boldsymbol{w}_1, \boldsymbol{x}_1, t_1) \propto p(t_1|\boldsymbol{x}_1, \boldsymbol{w}_1)\,\underbrace{p(t_0|\boldsymbol{x}_0, \boldsymbol{w}_0)p(\boldsymbol{w}_0)}_{p(\boldsymbol{w_1})} \tag{74}$$

After third observation:

$$p(\boldsymbol{w}_3|\boldsymbol{w}_2, \boldsymbol{x}_2, t_2) \propto p(t_2|\boldsymbol{x}_2, \boldsymbol{w}_2)\,\underbrace{p(t_1|\boldsymbol{x}_1, \boldsymbol{w}_1)\,\underbrace{p(t_0|\boldsymbol{x}_0, \boldsymbol{w}_0)p(\boldsymbol{w}_0)}_{p(\boldsymbol{w_1})}}_{p(\boldsymbol{w_2})} \tag{75}$$

Online linear regression

- Model: $t_i = \boldsymbol{w}_i^T \boldsymbol{x}_i + \epsilon_i$
- Likelihood: $p(t_i|\boldsymbol{x}_i, \boldsymbol{w}_i) = \mathcal{N}(t_i|\boldsymbol{w}_i^T \boldsymbol{x}_i, \beta^{-1})$
- Precision: $\beta = \frac{1}{\sigma^2}$
- Prior: $p(\boldsymbol{w}_0) = \mathcal{N}(\boldsymbol{w}_0|\boldsymbol{m}_0, \boldsymbol{S}_0)$
- Mean: $\boldsymbol{m}_0 = [0, \ldots, 0]$
- Covariance: $\boldsymbol{S}_0 = \alpha^{-1}\boldsymbol{I}$

Online linear regression - posterior

$$p(\boldsymbol{w}_{i+1}|\boldsymbol{w}_i, \boldsymbol{x}_i, t_i) = \mathcal{N}(\boldsymbol{w}_{i+1}|\boldsymbol{m}_{i+1}, \boldsymbol{S}_{i+1}) \tag{76}$$

$$\boldsymbol{S}_{i+1} = (\boldsymbol{S}_i^{-1} + \beta \boldsymbol{x}_i \boldsymbol{x}_i^T)^{-1} \tag{77}$$

$$\boldsymbol{m}_{i+1} = \boldsymbol{S}_{i+1}(\boldsymbol{S}_i^{-1}\boldsymbol{m}_i + \beta t_i \boldsymbol{x}_i) \tag{78}$$

Online linear regression - predictive

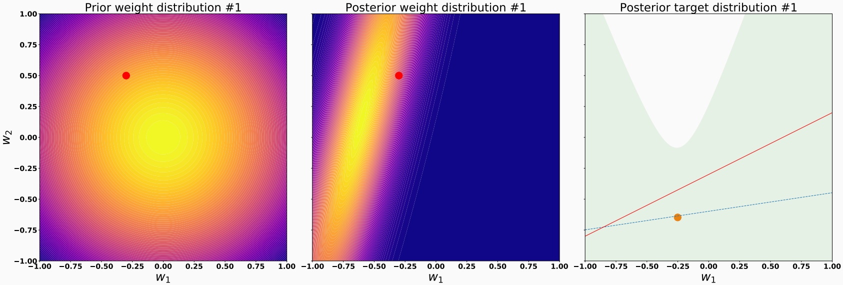
$$p(t_i|\boldsymbol{x}_i) = \mathcal{N}(t_i|\mu_i, \sigma_i) \tag{79}$$

$$\mu_i = \boldsymbol{w}_i^T \boldsymbol{x}_i \tag{80}$$

$$\sigma_i = \frac{1}{\beta} + \boldsymbol{x}_i^T \boldsymbol{S}_i \boldsymbol{x}_i \tag{81}$$

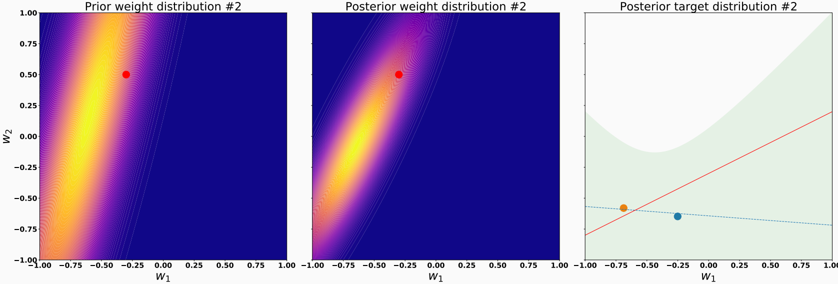
Online linear regression - progressive validation

$$y = w_1 + w_2 x,\ t = -0.3 + 0.5x + \epsilon, \beta = 25, \alpha = 2$$



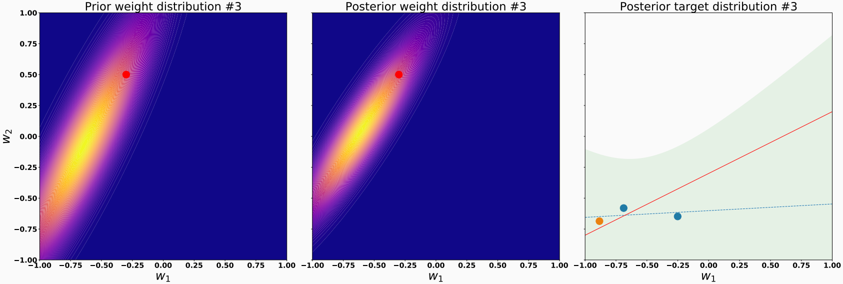
Online linear regression - progressive validation

$$y = w_1 + w_2 x,\ t = -0.3 + 0.5x + \epsilon, \beta = 25, \alpha = 2$$



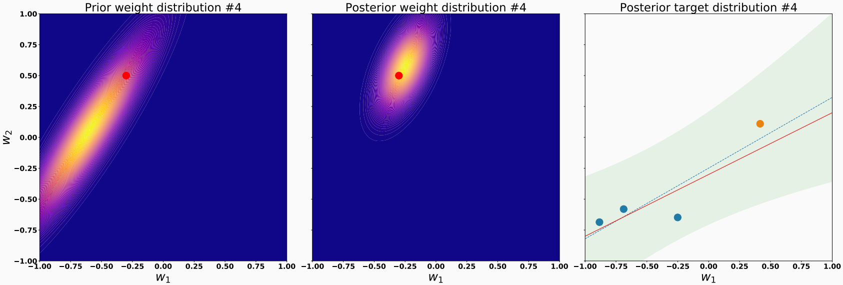
Online linear regression - progressive validation

$$y = w_1 + w_2 x,\ t = -0.3 + 0.5x + \epsilon, \beta = 25, \alpha = 2$$



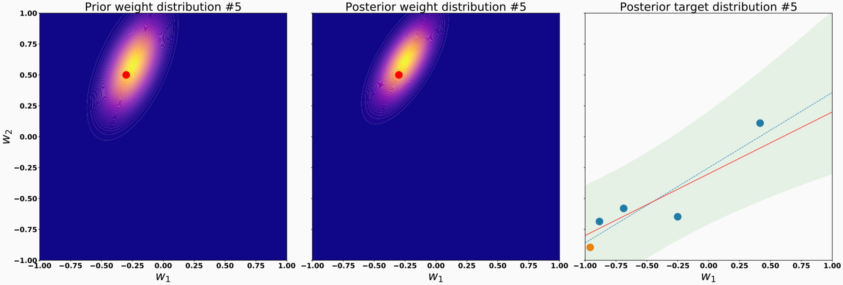
Online linear regression - progressive validation

$$y = w_1 + w_2 x,\ t = -0.3 + 0.5x + \epsilon, \beta = 25, \alpha = 2$$



Online linear regression - progressive validation

$$y = w_1 + w_2 x,\ t = -0.3 + 0.5x + \epsilon, \beta = 25, \alpha = 2$$



References

C. M. Bishop.
***Pattern Recognition and Machine Learning*.**
Springer, 2006.