Machine Learning

Lecture 11

Intro to BML

Anna Kuzina

Learning Outcomes

After this lecture you should know:

- What is the motivation for using bayesian approach
- What is prior and posterior
- Difference between MAP and MLE
- For Bayesian Linear Regression:
  - Derive posterior, MAP, MLE and predictive distribution

Background Joint and Marginal Distribution

Consider 2 random variables: x, y

Join distribution $p(x, y)$ (pdf) defines probabilities of all possibles pairs of x and y.

$$\int p(x, y) dx dy = 1 \quad \text{or} \quad \sum_x \sum_y p(x, y) = 1$$

Background Joint and Marginal Distribution

Consider 2 random variables: x, y

Join distribution $p(x, y)$ (pdf) defines probabilities of all possibles pairs of x and y.

$$\int p(x, y)dxdy = 1 \quad \text{or} \quad \sum_x \sum_y p(x, y) = 1$$

Marginal - pdf of a *single* r.v. (e.g. x) obtained from the joint distribution (sum rule)

Background Joint and Marginal Distribution

Consider 2 random variables: x, y

Join distribution $p(x, y)$ (pdf) defines probabilities of all possibles pairs of x and y.

$$\int p(x, y) dx dy = 1 \quad \text{or} \quad \sum_x \sum_y p(x, y) = 1$$

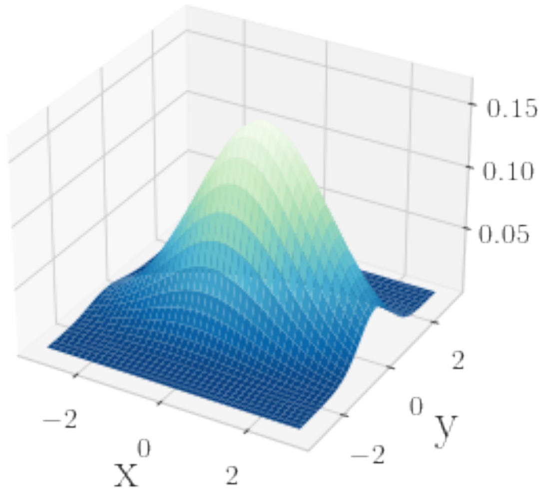
Marginal - pdf of a *single* r.v. (e.g. x) obtained from the joint distribution (sum rule)

$$p(x) = \int p(x, y) dy \quad \text{or} \quad p(x) = \sum_y p(x, y)$$

Background Joint and Marginal Distribution

Gaussian distribution $p(x, y) = \mathcal{N}((x, y) \mid \mu, \Sigma)$

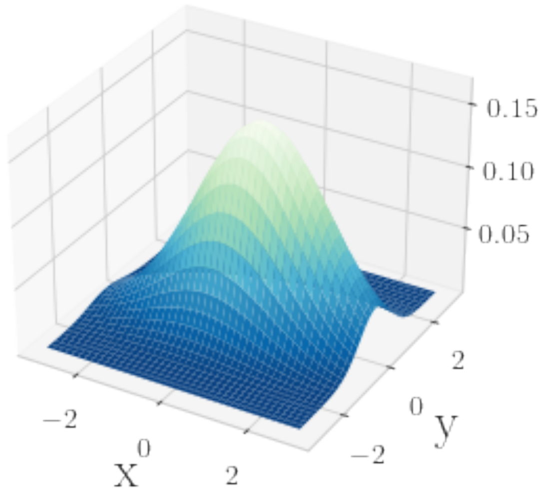
with $\mu = (\mu_x, \mu_y)$ and $\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$



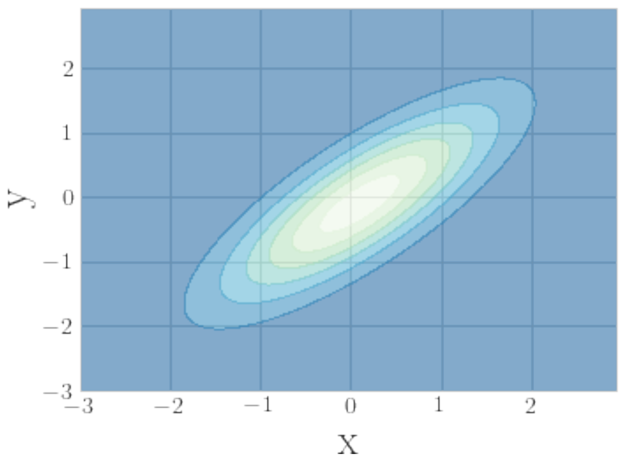
Background Joint and Marginal Distribution

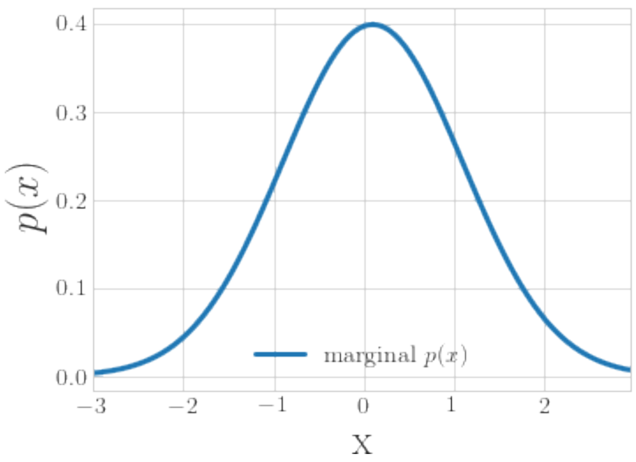
Gaussian distribution $p(x, y) = \mathcal{N}((x, y) \,|\, \mu, \Sigma)$

with $\mu = (\mu_x, \mu_y)$ and $\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$



Marginal $p(x) = \mathcal{N}(x \,|\, \mu_x, \sigma_{xx})$





Background Conditional Distribution

Conditional - probability of x given y (or distribution of x if we observe y)

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

Background Conditional Distribution

Conditional - probability of x given y (or distribution of x if we observe y)

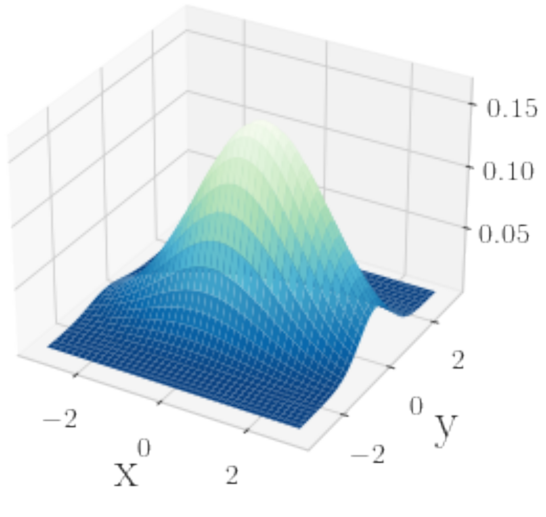
$$p(x|y) = \frac{p(x, y)}{p(y)}$$

Follows directly from the Product Rule

$$p(x, y) = p(x|y)p(y)$$

Background Joint and Marginal Distribution

Gaussian distribution $p(x, y) = \mathcal{N}((x, y) \,|\, \mu, \Sigma)$



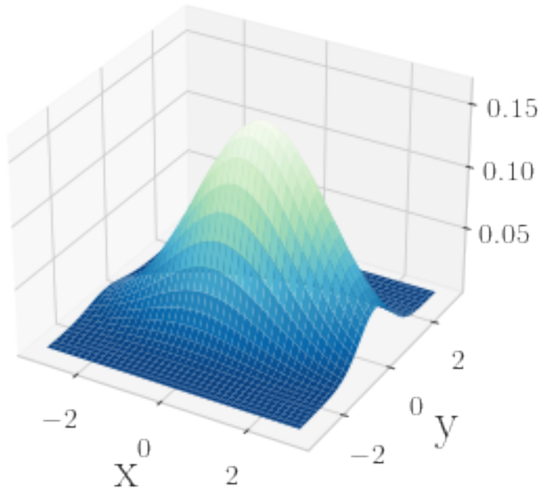
with $\mu = (\mu_x, \mu_y)$ and $\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$

Conditional $p(x \,|\, y = 1)$

Background Joint and Marginal Distribution

Gaussian distribution $p(x, y) = \mathcal{N}((x, y) \,|\, \mu, \Sigma)$

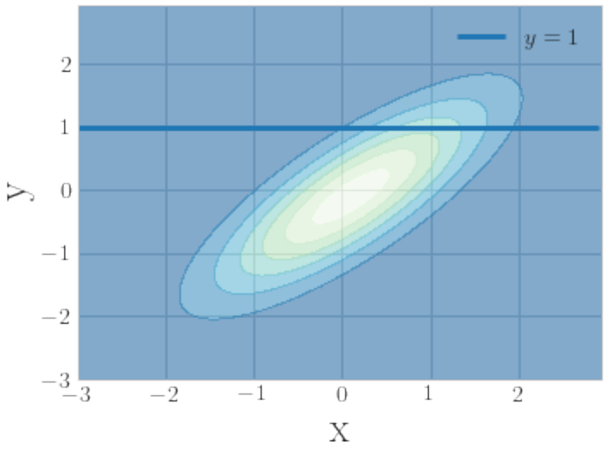
with $\mu = (\mu_x, \mu_y)$ and $\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$

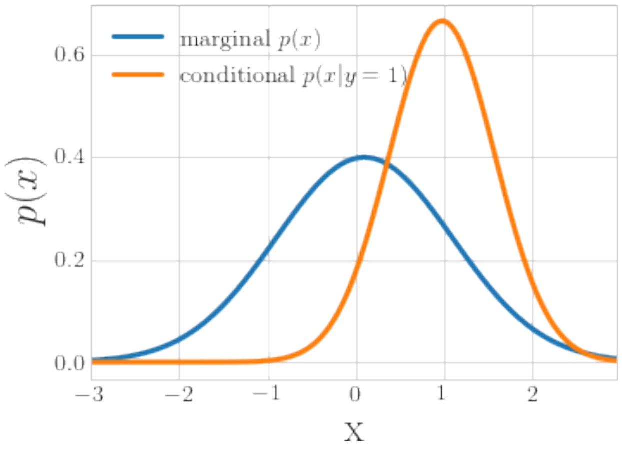


Conditional $p(x \,|\, y = 1) = \mathcal{N}(x \,|\, \bar{\mu}, \bar{\sigma})$

$\bar{\mu} = \mu_x + \sigma_{xy}\sigma_{yy}^{-1}(y - \mu_y)$

$\bar{\sigma} = \sigma_{xx} - \sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx}$





Background Example

sum rule

$$p(x) = \int p(x, y)dy$$

product rule

$$p(x, y) = p(x \mid y)p(y)$$

Example

Given $p(x, y, z)$ compute $p(y \mid x)$

Background Example

sum rule

$$p(x) = \int p(x, y) dy$$

product rule

$$p(x, y) = p(x \mid y) p(y)$$

Example

Given $p(x, y, z)$ compute $p(y \mid x)$

$$p(y \mid x) = \frac{p(y, x)}{p(x)} = \frac{\int p(x, y, z) dz}{\int p(x, y, z) dz dy}$$

Background Bayes Theorem

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

Background Bayes Theorem

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

Consider $\mathscr{D}$ - data that you observe, θ - unknown parameters

$$p(\theta|\mathscr{D}) = \frac{p(\mathscr{D}|\theta)p(\theta)}{p(\mathscr{D})} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} = \text{Posterior}$$

Background Statistical Inference

Observe: i.i.d. samples $x_1, \ldots, x_N$ from distribution $p(x \mid \theta)$

Goal: infer information about θ

- Frequentist approach (Maximum Likelihood)

- Bayesian Approach

Background Statistical Inference

Observe: i.i.d. samples $x_1, \ldots, x_N$ from distribution $p(x \mid \theta)$

Goal: infer information about θ

- Frequentist approach (Maximum Likelihood)

$$\theta^* = \arg\max_{\theta} p(X \mid \theta) = \arg\max_{\theta} \sum_i \log p(x_i \mid \theta)$$

- Bayesian Approach

$$p(\theta \mid x_1, \ldots, x_N) = \frac{\prod_i p(x_i \mid \theta) p(\theta)}{\int \prod_i p(x_i \mid \theta) p(\theta) d\theta}$$

Motivation Advantages of Bayesian Approach

- Encode prior knowledge or desired properties
- Prior imposes regularisation
- Posterior provides uncertainty about unknown parameters
- Bayesian ensembling

Models Discriminative and Generative

Data:

$\{x_1, \ldots, x_n\}$ - observed variables

$\{y_1, \ldots, y_N\}$ - unobserved / target variables

Model with unknown parameters θ:

- Discriminative

$$p(y, \theta | x) = p(y | x, \theta)p(\theta)$$

- Generative

$$p(x, y, \theta) = p(x, y | \theta)p(\theta)$$

Models Training Time

Input:

- Dataset $\{x_n, y_n\}_{n=1}^{N}$
- Likelihood $p(y \mid x, \theta)$
- Prior $p(\theta)$

Output:

Posterior distribution

$$p(\theta \mid X, Y) = \frac{\prod_n p(y_n \mid x_n, \theta) p(\theta)}{\int \prod_n p(y_n \mid x_n, \theta) p(\theta) d\theta}$$

Models Testing Time

Input:

Posterior distribution $$p(\theta | X, Y) = \frac{\prod_n p(y_n | x_n, \theta) p(\theta)}{\int \prod_n p(y_n | x_n, \theta) p(\theta) d\theta}$$

Test point: x^*

Output:

Predictive Distribution

$$p(y^* | x^*, \theta, X, Y) = \int p(y^* | x^*, \theta) p(\theta | X, Y) d\theta$$

Models: How to find posterior?

$$p(\theta | X, Y) = \frac{\prod_n p(y_n | x_n, \theta) p(\theta)}{\int \prod_n p(y_n | x_n, \theta) p(\theta) d\theta}$$

- Conjugacy

  For ‘good’ pairs of likelihood and prior

- MAP estimate (maximum a posteriori)

  Point estimate instead of distribution

- Approximate Inference

  Find something that looks like posterior, have samples like posterior

Linear Regression: Recap

Training Dataset: $X \in \mathbb{R}^{N \times d}$; $\quad Y \in \mathbb{R}^N$

Model: $a(X) = Xw$

Cost function: $\mathscr{L}(a, X) = \frac{1}{N}\|Xw - Y\|^2$

$$w^* = (X^T X)^{-1} X^T Y$$

If we add regularization

$$w^* = (\lambda I + X^T X)^{-1} X^T Y$$

Linear Regression: Additive Noise

Training Dataset: $X \in \mathbb{R}^{N \times d}; \quad Y \in \mathbb{R}^N$

Model:

$p(y \mid w, x) = \mathcal{N}(y \mid x^T w, \sigma^2)$ - likelihood

$p(w) = \mathcal{N}(w \mid 0, A^{-1})$ - prior

Recap Gaussian Distribution

Recap pdf of gaussian distribution

Recap Gaussian Distribution

Recap pdf of gaussian distribution

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

We will be working with logarithms a lot:

Recap Gaussian Distribution

Recap pdf of gaussian distribution

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

We will be working with logarithms a lot:

$$\log \mathcal{N}(x \mid \mu, \Sigma) = -\frac{1}{2}(x-\mu)^T\Sigma(x-\mu) + const$$

It is **always** true that:

μ is mode of the distribution: $\mu = \arg\max_x \log \mathcal{N}(x \mid \mu, \Sigma)$

Σ is the inverse Hessian: $\Sigma = -\left(\nabla_x^2 \log \mathcal{N}(x \mid \mu, \Sigma)\right)^{-1}$

Linear Regression: MLE and MAP

Let's start with MLE:

$$p(Y|w, X) = \prod_{n=1}^{N} p(y_n|w, x_n)$$

$$w^{MLE} = \arg\max_{w} \sum_{n} \log \mathcal{N}(y_n|w, x_n)$$

Linear Regression: MLE and MAP

Let's start with MLE:

$$p(Y|w,X) = \prod_{n=1}^{N} p(y_n|w,x_n)$$

$$w^{MLE} = \arg\max_w \sum_n \log \mathcal{N}(y_n|w,x_n) = (X^T X)^{-1} X^T Y$$

Linear Regression: MLE and MAP

Now, MAP: $p(Y|w,X) = \prod_{n=1}^{N} p(y_n|w,x_n), \quad p(w) = \mathcal{N}(w|0,A^{-1})$

Maximize Posterior $p(w|X,Y) = \dfrac{p(Y|w,X)p(w)}{\int p(Y|w,X)p(w)dw}$

Linear Regression: MLE and MAP

Now, MAP: $p(Y|w,X) = \prod_{n=1}^{N} p(y_n|w,x_n), \quad p(w) = \mathcal{N}(w|0,A^{-1})$

Maximize Posterior $p(w|X,Y) = \dfrac{p(Y|w,X)p(w)}{\int p(Y|w,X)p(w)dw}$

$$w^{MAP} = (A + \frac{1}{\sigma^2}X^TX)^{-1}\frac{1}{\sigma^2}X^TY$$

Linear Regression: MLE and MAP

Now, MAP: $p(Y|w, X) = \prod_{n=1}^{N} p(y_n|w, x_n), \quad p(w) = \mathcal{N}(w|0, A^{-1})$

Maximize Posterior $p(w|X, Y) = \dfrac{p(Y|w, X)p(w)}{\int p(Y|w, X)p(w)dw}$

$$w^{MAP} = (A + \frac{1}{\sigma^2}X^TX)^{-1}\frac{1}{\sigma^2}X^TY$$
$$w^{MLE} = (X^TX)^{-1}X^TY$$

Linear Regression: Posterior

$$p(Y \mid w, X) = \prod_{n=1}^{N} p(y_n \mid w, x_n), \quad p(w) = \mathcal{N}(w \mid 0, A^{-1})$$

$$p(w \mid X, Y) = \frac{p(Y \mid w, X) p(w)}{\int p(Y \mid w, X) p(w) dw}$$

Linear Regression: Posterior

$$p(Y|w,X) = \prod_{n=1}^{N} p(y_n|w,x_n), \quad p(w) = \mathcal{N}(w|0,A^{-1})$$

$$p(w|X,Y) = \frac{p(Y|w,X)p(w)}{\int p(Y|w,X)p(w)dw} = \mathcal{N}(w|w^{MAP},(A+\frac{1}{\sigma^2}X^TX)^{-1})$$

Linear Regression: Predictive Distribution

Given: posterior distribution $p(w \mid X, Y)$

New point: x^*

Predictive distribution:

Linear Regression: Predictive Distribution

Given: posterior distribution $p(w \mid X, Y)$

New point: $x*$

Predictive distribution:

$$p(y* \mid x*, w, X, Y) = \int p(y* \mid x*, w) p(w \mid X, Y) dw = \mathcal{N}(y* \mid x^{*T} \mu_w, \sigma^2 + x^{*T} \Sigma_w x*)$$

Linear Regression: Hyperparameters

Where to get unknown A and σ^2?

Evidence maximization.

Linear Regression: Hyperparameters

Where to get unknown A and σ^2?

Evidence maximization. Consider hyperpameter $\alpha = \{A, \sigma^2\}$.

Impose uninformative prior $p(\alpha) \approx$ const. Then, MAP of the hyperparameter is:

Linear Regression: Hyperparameters

Where to get unknown A and σ^2?

Evidence maximization. Consider hyperpameter $\alpha = \{A, \sigma^2\}$.

Impose uninformative prior $p(\alpha) \approx$ const. Then, MAP of the hyperparameter is:

$$\arg\max_{\alpha} p(\alpha | X, Y) = \arg\max_{\alpha} \frac{p(Y | X, \alpha) p(\alpha)}{p(Y | X)} = \arg\max_{\alpha} p(Y | X, \alpha)$$

That is, we need to maximise **evidence** to get optimal hyperparameters.

Learning Outcomes

After this lecture you should know:

- What is the motivation for using bayesian approach
- What is prior and posterior
- Difference between MAP and MLE
- For Bayesian Linear Regression:
  - Derive posterior, MAP, MLE and predictive distribution