Machine Learning

CMPT 726

Mo Chen
SFU School of Computing Science

2022-10-11

Probabilistic Interpretation of Linear Regression

Probabilistic Model

Suppose we know how the data was generated (known as the “data generating process”):

$y = \vec{w}^\top \vec{x} + \sigma\epsilon,$ where $\epsilon \sim \mathcal{N}(0,1)$

$\vec{w}$ and σ are unknown parameters

We observe many tuples $(\vec{x}, y)$, denoted as $\{(\vec{x}_i, y_i)\}_{i=1}^{N} =: \mathcal{D}$, which we assume are generated i.i.d. from the process above.

Terminology: i.i.d. stands for “independent and identically distributed”, which means that each tuple $(\vec{x}_i, y_i)$ is independent of other tuples and has the same joint distribution as any other tuple.

Goal: Estimate the values of unknown parameters from observations (known as “parameter estimation”).

Probabilistic Model

A **probabilistic model** assigns a probability to every possible observation.

We choose a probabilistic model based on the data generating process:

Recall: data generating process: $y = \vec{w}^\top \vec{x} + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$

Recall: if $Z \sim \mathcal{N}(0,1)$, $\mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$

So, $y|\vec{x}, \vec{w}, \sigma \sim \mathcal{N}(\vec{w}^\top \vec{x}, \sigma^2)$

We can write down the expression for $p(y|\vec{x}, \vec{w}, \sigma)$:

Recall: If $X \sim \mathcal{N}(\mu, \sigma^2)$, $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$$p(y|\vec{x}, \vec{w}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \vec{w}^\top \vec{x})^2}{2\sigma^2}\right)$$

Parameter Estimation

We have two unknown parameters: $\vec{w}$ and σ.

For the purposes of making predictions, we care mostly about $\vec{w}$ and so will focus on estimating $\vec{w}$.

In general:

The parameters we care about are known as **parameters of interest**.

The parameters we don’t care about are known as **nuisance parameters**.

In this case, $\vec{w}$ is the parameter of interest, and σ is the nuisance parameter.

Maximum Likelihood Estimation (MLE)

Idea: Find the parameter value at which the probability of observing the data is maximized.

Step 1: Derive the joint probability density of the observations $p(\mathcal{D}|\vec{w},\sigma)$.

$$p(\mathcal{D}|\vec{w},\sigma) := p(y_1,\ldots,y_N|\vec{x}_1,\ldots,\vec{x}_N,\vec{w},\sigma)$$

$$= \prod_{i=1}^{N} p(y_i|\vec{x}_1,\ldots,\vec{x}_N,\vec{w},\sigma)$$ (Conditioned on $\vec{x}_1,\ldots,\vec{x}_N$, the y_i's are independent)

$$= \prod_{i=1}^{N} p(y_i|\vec{x}_i,\vec{w},\sigma)$$ (Conditioned on each $\vec{x}_i$, y_i and $\vec{x}_j$'s for $j \neq i$ are independent)

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\vec{w}^\top\vec{x}_i)^2}{2\sigma^2}\right)$$

$$:= \mathcal{L}\left(\vec{w},\sigma;\{(\vec{x}_i,y_i)\}_{i=1}^{N}\right) = \mathcal{L}(\vec{w},\sigma;\mathcal{D})$$

“Likelihood function”

Maximum Likelihood Estimation (MLE)

Step 2: Find the value of the parameter of interest that maximizes the likelihood function.

$$\widehat{\vec{w}}_{\mathrm{MLE}} = \arg\max_{\vec{w}} \mathcal{L}(\vec{w}, \sigma; \mathcal{D})$$

To find it, we can try computing the gradient and set it to zero.

$$\frac{\partial}{\partial \vec{w}} \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) = \frac{\partial}{\partial \vec{w}} \left(\prod_{i=1}^{N} p(y_i | \vec{x}_i, \vec{w}, \sigma) \right)$$

$$= \sum_{i=1}^{N} \left(\frac{\partial}{\partial \vec{w}} p(y_i | \vec{x}_i, \vec{w}, \sigma) \prod_{j \neq i} p(y_j | \vec{x}_j, \vec{w}, \sigma) \right)$$

Unwieldy!

Maximum Likelihood Estimation (MLE)

Step 2: Find the value of the parameter of interest that maximizes the likelihood function.

$$\widehat{\vec{w}}_{\mathrm{MLE}} = \arg\max_{\vec{w}} \mathcal{L}(\vec{w}, \sigma; \mathcal{D})$$

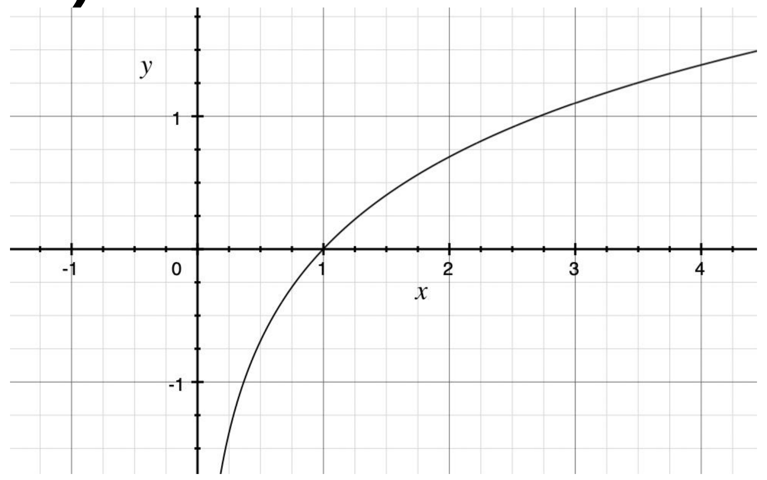
When the base of the log is not shown, it defaults to e (natural number)

Instead use the fact that $\widehat{\vec{w}}_{\mathrm{MLE}} = \arg\max_{\vec{w}} \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) = \arg\max_{\vec{w}} \log \mathcal{L}(\vec{w}, \sigma; \mathcal{D})$.

(This is true because $x \mapsto \log(x)$ is strictly increasing, i.e.: as x increases, $\log(x)$ increases)

$$\log \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) = \log \prod_{i=1}^{N} p(y_i | \vec{x}_i, \vec{w}, \sigma) = \sum_{i=1}^{N} \log p(y_i | \vec{x}_i, \vec{w}, \sigma)$$

"Log-likelihood function"



This turns the expression into a sum, which is easy to differentiate.

$$\frac{\partial}{\partial \vec{w}} \log \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) = \frac{\partial}{\partial \vec{w}} \left(\sum_{i=1}^{N} \log p(y_i | \vec{x}_i, \vec{w}, \sigma) \right) = \sum_{i=1}^{N} \frac{\partial}{\partial \vec{w}} \log p(y_i | \vec{x}_i, \vec{w}, \sigma)$$

Maximum Likelihood Estimation (MLE)

Step 2: Find the value of the parameter of interest that maximizes the likelihood function.

$$\widehat{\vec{w}}_{\mathrm{MLE}} = \arg\max_{\vec{w}} \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) = \arg\max_{\vec{w}} \log \mathcal{L}(\vec{w}, \sigma; \mathcal{D})$$

$$\log \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) = \sum_{i=1}^{N} \log p(y_i | \vec{x}_i, \vec{w}, \sigma)$$

$$= \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \vec{w}^\top \vec{x}_i)^2}{2\sigma^2} \right) \right)$$

$$= \sum_{i=1}^{N} \left(-\frac{(y_i - \vec{w}^\top \vec{x}_i)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2} \right)$$

$$= -\sum_{i=1}^{N} \frac{(y_i - \vec{w}^\top \vec{x}_i)^2}{2\sigma^2} - N\log\sqrt{2\pi\sigma^2}$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 - N\log\sqrt{2\pi\sigma^2}$$

Maximum Likelihood Estimation (MLE)

$$\begin{aligned}
\widehat{\vec{w}}_{\text{MLE}} &= \underset{\vec{w}}{\operatorname{argmax}} \log \mathcal{L}(\vec{w}, \sigma; \mathcal{D}) \\
&= \arg\max_{\vec{w}} \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 - N \log \sqrt{2\pi\sigma^2} \right) \\
&= \arg\min_{\vec{w}} \left(\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 + N \log \sqrt{2\pi\sigma^2} \right) \\
&= \arg\min_{\vec{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 \\
&= \arg\min_{\vec{w}} \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2
\end{aligned}$$

Compare to the loss function of OLS:

$$L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2$$

Frequentist vs. Bayesian Statistics

Previously we treated parameters as *fixed* (albeit unknown) quantities.

Suppose we know what values of the parameters are more plausible compared to others. This is known as **prior belief** or **prior knowledge**.

We can represent this prior belief as a probability distribution over parameters, which is known as a **prior probability distribution**.

Instead of treating parameters as fixed quantities, we treat the parameters as random variables that follow the prior probability distribution.

This latter approach is known as **Bayesian statistics**, whereas the former approach is known as **frequentist statistics**.

Maximum A Posteriori Estimation (MAP)

In addition to the probabilistic model, we have a prior distribution $p(\vec{\theta})$ over the parameters of interest $\vec{\theta}$.

Idea: Find the parameter value that maximizes the conditional probability over parameters given the observations $p\left(\vec{\theta}\middle|\mathcal{D}\right)$.

We can use Bayes' rule to find $p\left(\vec{\theta}\middle|\mathcal{D}\right)$:

“Prior” “Likelihood”

“Posterior”

$$p\left(\vec{\theta}\middle|\mathcal{D}\right) = \frac{p\left(\vec{\theta}\right)p\left(\mathcal{D}\middle|\vec{\theta}\right)}{p(\mathcal{D})}$$

$$\hat{\vec{\theta}}_{\text{MAP}} = \arg\max_{\vec{\theta}} p\left(\vec{\theta}\middle|\mathcal{D}\right)$$

The term “likelihood” is used in broader contexts than “likelihood function”. While its meaning coincides with “likelihood function” in this context, it could mean the probability of a single observation $p\left(y\middle|\vec{x},\vec{\theta}\right)$ or any $p(\vec{x}|\vec{y})$ where we observe $\vec{x}$ and would like to infer an unobserved $\vec{y}$ based on $\vec{x}$.

Maximum A Posteriori Estimation (MAP)

Consider the same probabilistic model as before:

$$y|\vec{x},\vec{w},\sigma \sim \mathcal{N}(\vec{w}^\top\vec{x},\sigma^2)$$

$$p(y|\vec{x},\vec{w},\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\vec{w}^\top\vec{x})^2}{2\sigma^2}\right)$$

We choose the following prior distribution over the parameter of interest $\vec{w}$:

$$\vec{w}|\sigma \sim \mathcal{N}\left(\vec{0},\frac{\sigma^2}{\lambda}I\right)$$

$$p(\vec{w}|\sigma) = \frac{1}{\sqrt{(2\pi)^n \det\left(\frac{\sigma^2}{\lambda}I\right)}}\exp\left(-\frac{1}{2}\vec{w}^\top\left(\frac{\sigma^2}{\lambda}I\right)^{-1}\vec{w}\right) = \frac{1}{\sqrt{\left(\frac{2\pi\sigma^2}{\lambda}\right)^n}}\exp\left(-\frac{\lambda}{2\sigma^2}\|\vec{w}\|_2^2\right)$$

Maximum A Posteriori Estimation (MAP)

$$
\begin{aligned}
\widehat{\vec{w}}_{\mathrm{MAP}} &= \arg\max_{\vec{w}} p(\vec{w}|\mathcal{D},\sigma) \\
&= \arg\max_{\vec{w}} \log p(\vec{w}|\mathcal{D},\sigma) \\
&= \arg\max_{\vec{w}} \log\left(\frac{p(\vec{w}|\sigma)p(\mathcal{D}|\vec{w},\sigma)}{p(\mathcal{D}|\sigma)}\right) \\
&= \arg\max_{\vec{w}} [\log p(\vec{w}|\sigma) + \log p(\mathcal{D}|\vec{w},\sigma) - \log p(\mathcal{D}|\sigma)] \\
&= \arg\max_{\vec{w}} [\log p(\vec{w}|\sigma) + \log p(\mathcal{D}|\vec{w},\sigma)] \\
&= \arg\max_{\vec{w}} [\log p(\vec{w}|\sigma) + \log p(y_1,\dots,y_N|\vec{x}_1,\dots,\vec{x}_N,\vec{w},\sigma)] \\
&= \arg\max_{\vec{w}} \left[\log p(\vec{w}|\sigma) + \log \prod_{i=1}^{N} p(y_i|\vec{x}_i,\vec{w},\sigma)\right] \\
&= \arg\max_{\vec{w}} \left[\log p(\vec{w}|\sigma) + \sum_{i=1}^{N} \log p(y_i|\vec{x}_i,\vec{w},\sigma)\right]
\end{aligned}
$$

Maximum A Posteriori Estimation (MAP)

$$
\begin{aligned}
\widehat{\vec{w}}_{\mathrm{MAP}} &= \arg\max_{\vec{w}}[\log p(\vec{w}|\sigma) + \sum_{i=1}^{N} \log p(y_i|\vec{x}_i, \vec{w}, \sigma)] \\
&= \arg\max_{\vec{w}}\left[\log\left(\frac{1}{\sqrt{\left(\frac{2\pi\sigma^2}{\lambda}\right)^n}}\exp\left(-\frac{\lambda}{2\sigma^2}\|\vec{w}\|_2^2\right)\right) + \sum_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \vec{w}^\top\vec{x}_i)^2}{2\sigma^2}\right)\right)\right] \\
&= \arg\max_{\vec{w}}\left[-\frac{\lambda}{2\sigma^2}\|\vec{w}\|_2^2 - \log\left(\sqrt{\left(\frac{2\pi\sigma^2}{\lambda}\right)^n}\right) + \sum_{i=1}^{N}\left(-\frac{(y_i - \vec{w}^\top\vec{x}_i)^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right)\right] \\
&= \arg\max_{\vec{w}}\left[-\frac{\lambda}{2\sigma^2}\|\vec{w}\|_2^2 + \sum_{i=1}^{N}\left(-\frac{(y_i - \vec{w}^\top\vec{x}_i)^2}{2\sigma^2}\right)\right]
\end{aligned}
$$

Maximum A Posteriori Estimation (MAP)

$$\hat{\vec{w}}_{\text{MAP}} = \arg\max_{\vec{w}} \left[-\frac{\lambda}{2\sigma^2} \|\vec{w}\|_2^2 + \sum_{i=1}^{N} \left(-\frac{(y_i - \vec{w}^\top \vec{x}_i)^2}{2\sigma^2} \right) \right]$$

$$= \arg\max_{\vec{w}} - \frac{1}{2\sigma^2} \left(\lambda \|\vec{w}\|_2^2 + \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 \right)$$

$$= \arg\min_{\vec{w}} \frac{1}{2\sigma^2} \left(\lambda \|\vec{w}\|_2^2 + \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 \right)$$

$$= \arg\min_{\vec{w}} \left(\lambda \|\vec{w}\|_2^2 + \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 \right)$$

Compare to the loss function of ridge regression: $L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 + \lambda \|\vec{w}\|_2^2$

OLS vs. Ridge Regression (Deterministic Interpretation)

Model: $\vec{y} = \vec{w}^{\top}\vec{x}$

Parameters: $\vec{w}$

OLS:

Loss function:

$$L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^{\top}\vec{x}_i)^2$$

$$= \|\vec{y} - X\vec{w}\|_2^2$$

Optimal Parameters:

$$\vec{w}^{*} \coloneqq \arg\min_{\vec{w}} L(\vec{w}) = (X^{\top}X)^{-1}X^{\top}\vec{y}$$

Ridge Regression:

Loss function:

$$L(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^{\top}\vec{x}_i)^2 + \lambda\|\vec{w}\|_2^2$$

$$= \|\vec{y} - X\vec{w}\|_2^2 + \lambda\left\|\vec{w}\right\|_2^2$$

Optimal Parameters:

$$\vec{w}^{*}: = \arg\min_{\vec{w}} L(\vec{w}) = (X^{\top}X + \lambda I)^{-1}X^{\top}\vec{y}$$

OLS vs. Ridge Regression (Probabilistic Interpretation)

Probabilistic Model: $y|\vec{x}, \vec{w}, \sigma \sim \mathcal{N}(\vec{w}^\top \vec{x}, \sigma^2)$; Prior: $\vec{w}|\sigma \sim \mathcal{N}\left(\vec{0}, \frac{\sigma^2}{\lambda} I\right)$

Parameters: $\vec{w}$

Nuisance Parameter: σ

MLE estimate:

$$\widehat{\vec{w}}_{\mathrm{MLE}} = \arg\min_{\vec{w}} \sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2$$

$$= \arg\min_{\vec{w}} \|\vec{y} - X\vec{w}\|_2^2$$

$$= (X^\top X)^{-1} X^\top \vec{y}$$

MAP estimate:

$$\widehat{\vec{w}}_{\mathrm{MAP}} = \arg\min_{\vec{w}} \left[\sum_{i=1}^{N} (y_i - \vec{w}^\top \vec{x}_i)^2 + \lambda \|\vec{w}\|_2^2\right]$$

$$= \arg\min_{\vec{w}} [\|\vec{y} - X\vec{w}\|_2^2 + \lambda \|\vec{w}\|_2^2]$$

$$= (X^\top X + \lambda I)^{-1} X^\top \vec{y}$$

Takeaways

There are deep connections between deterministic and probabilistic formulations of machine learning methods.

Many methods have both deterministic and probabilistic interpretations.

Primary loss function is related to the likelihood.

Regularizer is related to the prior.

Square losses are related to (univariate) Gaussian observation noise.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Squared L2 losses are related to (multivariate) isotropic Gaussian observation noise.

$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2} \|\vec{x}-\vec{\mu}\|_2^2\right)$$

(Full) Bayesian Inference

Bayesian inference refers to computing the full posterior:

$$p\left(\vec{\theta}\middle|\mathcal{D}\right) = \frac{p\left(\vec{\theta}\right)p\left(\mathcal{D}\middle|\vec{\theta}\right)}{p(\mathcal{D})}$$

Compare to MAP:

$$\hat{\vec{\theta}}_{\mathrm{MAP}} = \arg\max_{\vec{\theta}} p\left(\vec{\theta}\middle|\mathcal{D}\right)$$

Unlike MAP estimation, it does not just produce a single value for the parameter estimate (known as a **point estimate**). Instead, it produces a full distribution over possible parameter values.

Can be extremely computationally challenging - computing the posterior exactly is intractable for all but some special cases. Must typically rely on sampling or approximations.

Terminology: “Modes”

A **mode** of a probability distribution is:

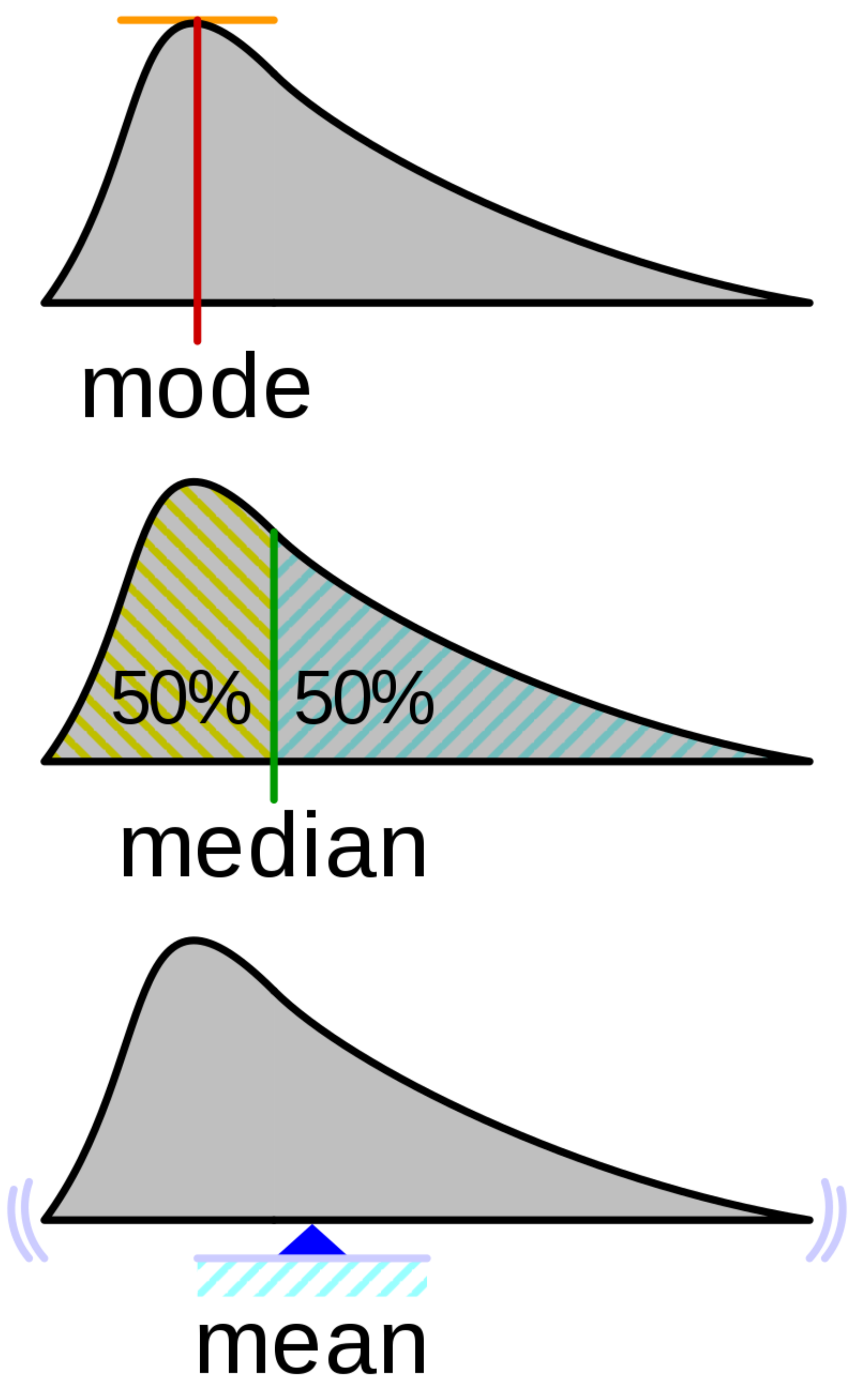
- in the case of continuous distribution: a local maximum of the probability density

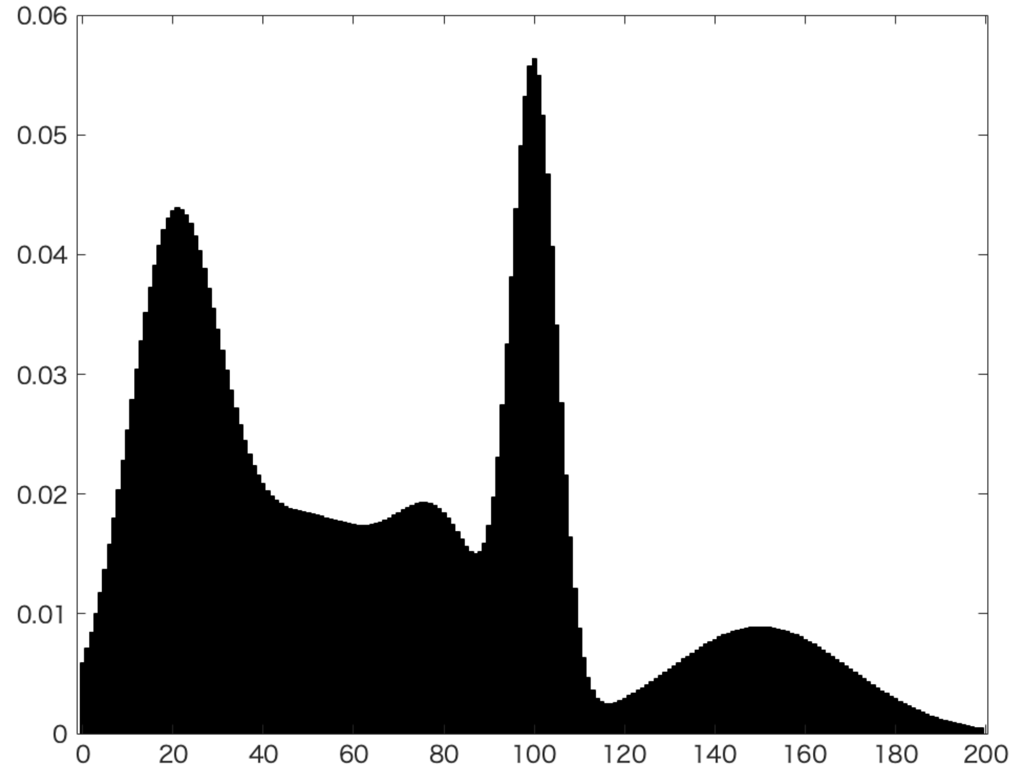
- In the case of discrete distribution: the values at which the probability mass is the highest

“Unimodal”: a probability distribution with a single mode (e.g.: Gaussian, Gamma, log-normal, etc.)

“Bimodal”: a probability distribution with two modes (e.g.: a mixture of two Gaussians)

“Multimodal”: a probability distribution with many modes (e.g.: a mixture of k Gaussians)





Credit: Wikipedia

Full Bayesian Inference vs. MAP

When the posterior $p\left(\vec{\theta}\middle|\mathcal{D}\right)$ is unimodal:

- There is one parameter value that is the most plausible given the data.
- The MAP estimate corresponds to this parameter value.
- The MAP estimate would capture the most important information of the posterior well.

When the posterior $p\left(\vec{\theta}\middle|\mathcal{D}\right)$ is multimodal:

- There could be multiple parameter values that are all quite plausible (or even equally plausible).
- The MAP estimate corresponds to one of these parameter values (in the case of multiple equally plausible values, which one it corresponds to is arbitrary).
- The MAP estimate would lose a lot of information in the posterior.

Note on the Term “Inference”

Parameter estimation is also known as **inference** (especially in the context of Bayesian inference), since the purpose is to infer the value of unknown parameters from observed data.

Note: In machine learning, the term “inference” is overloaded and can sometimes mean testing/making predictions (as opposed to training).

This can get confusing, because parameter estimation in a probabilistic model corresponds to training the model, which is also known as inference. On the other hand, testing the model is also sometimes known as inference. So, inference can mean either training or testing!

Meaning depends on context:

- “inference time”, “training and inference”: means testing
- “Bayesian inference”, “probabilistic inference”, “parameter inference”: means training
- “inference procedure”: ambiguous

In this course, we will avoid using the term as much as possible to minimize confusion.

Terminology: “A Priori” vs. “A Posteriori”

“A Priori” (literally means “from the earlier”): Prior knowledge that we have *before* seeing the data

“A Posteriori” (literally means “from the later”): Knowledge acquired *after* seeing the data

Examples:

“The degree of the polynomial is not known *a priori*” means that we don’t know what the degree of the polynomial should be without seeing the data.

“The *a posteriori* estimates are mostly concentrated around 0.5” means that after seeing the data, most of the estimates are near 0.5.

$$\overset{\text{“Posterior”}}{p(\vec{\theta}|\mathcal{D})} = \frac{\overset{\text{“Prior”}}{p(\vec{\theta})}\overset{\text{“Likelihood”}}{p(\mathcal{D}|\vec{\theta})}}{p(\mathcal{D})}$$