BEST VALUES FOR MODEL PARAMETERS __/__/____

(15) Maximum Likelihood Estimation

Parameter Estimation: We often assume a parametric model in ml or stats, ie data is generated from some distribution $P(x;\theta)$ or $P(y|x;\theta)$ governed by params θ

Goal: Estimate θ from observed data.

θ, that makes the observed data most likely

Likelihood function: $L(\theta) = P(x_1, x_2 .. x_n | \theta)$

If samples are independant: $L(\theta) = \prod_{i=1}^{n} P(x_i|\theta)$

MLE choose, $\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$

To simplify math, we take log. → $\hat{\theta}_{MLE} = \arg\max_{\theta} l(\theta)$

Why ? Why ?

- Every ML model has some parameters associated with them.
- We don't know their values initially
- So we must estimate them from training data.

In logistic regression → MLE: cross-entropy loss

In linear regression → MLE: MSE

MLE = pure data fitting (no prior assumptions)

* Depending on what probability distribution you assume for data you get different loss functions
* Every loss function in ML and DL is just the negative log likelihood of a suitable probabilistic model.

__/__/____

Example: Linear Regression

1. Assume: $y_i = w^T x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ {Normal}

2. Likelihood:

$$P(y_i | x_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y_i - w^T x_i}{2\sigma^2}\right)^2$$

3log likelihood:

$$l(w) = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + C$$

4. Negative log likelihood: $J(w) = \sum_i (y_i - w^T x_i)^2$

This is the frickin MSE!

MLE work great when you've
- Plenty of data
- No noise
- Parameters are identifiable

MLE tends to overfit data (training) because it only tries to maximize fit - it doesn't penalize large or extreme parameter values.

fix?

(15) Maximum A Posteriori Estimation. (MAP)

Now we take a Bayesian,

Instead of just maximising $P(D|\Theta)$ we also consider how likely the parameter itself is - using a prior distribution $P(\Theta)$

__/__/____

Bayes Rule: $P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$

Then: $\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|D)$

Since $P(D)$ is constant $\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(P(D|\theta)\,P(\theta)\right)$

$\approx \hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[\log P(D|\theta) + \log P(\theta)\right]$

↙ data fitting term (likelihood)

↘ regularization term

{ MAP = MLE + Regularization }

Gaussian Prior → L2 | MAP works well for small datas
Laplace Prior → L1 | uncertain params.

- MLE: defines the data fit part of the loss.
  MLE gives us the core cost / loss function of the model.
  → Best θ maximising likelihood
- MAP: adds a prior on parameters (regularization)
  MAP gives loss + regularization term.
  → Best θ maximising posterior.

⑯ Bayesian Thinking: is a way of doing inference - ie, learning about unknown quantities based on data using Bayes Theorem.

{Posterior = Prior × Likelihood}

__/__/____

→ Frequentist (MLE) : Fixed but unknown constraints.

- Single best θ (max likelihood)
- It outputs a point estimate

→ Bayesian (MAP) : Random variables with probability distributio

- Full distribution
- Uses a prior → output: updates belief.

{Use the data - vs - update myself If I'm wrong}

(17) Bayesian Estimation

Instead of finding a single best parameter value (like MAP / MLE)
we compute the entire probability distribution of the
parameters after seeing data.

So instead of a single $\hat{\theta}$, we have a full posterior dist
$P(\theta|D)$ → Bayesian estimation.

Bayesian Linear Regression, model = $y = w^T x + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

1. Starting with prior = $w \sim N(0, \tau^2 I)$

2. Compute likelihood = $P(D|w) \propto \exp\left(-\frac{1}{2\sigma^2} \|y - Xw^2\|^2\right)$

3. Posterior = $P(w|D) \propto \exp\left(-1/2\sigma^2 \|y - Xw\|^2\right) \exp\left(-1/2\tau^2 \|w\|^2\right)$
by Bayes Theoram

This posterior is also gaussian.

4. We take log and negate: $\text{loss}(w) = \|y - Xw\|^2 + \lambda \|w\|^2$

≈ Ridge Regression - the MAP estimate of the Bayesian Model.

__/__/____

In easy terms ⇒
- Prior: What you believe before data
- Likelihood: How data pushes your belief
- Posterior: Your updated belief
- MAP: Most likely belief, after update

Bayesian Estimation: Full range of beliefs after update.

⇒ Why use Bayesian Estimation
- Uncertainty estimation
- Small data robustness
- Regularization built-in
- Sequential Updating

(18) Complete Distributions

A probability distribution describes how likely different outcomes are. Tells how randomness behaves.

→ Discrete distributions: events that take specific values

① Bernoulli distribution: Used for binary outcomes

$$P(X = x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

→ logistic regression: Binary classification

② Binomial Distribution: Sum of n Bernoulli trials

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

→ Continuous distribution: events that take any real value

① Normal / Gaussian distribution: classic bell curve

$$P(x) = 1/\sqrt{2\pi\sigma^2}\; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$