## 1. Conditional probability

$$P(A \cap B) = P(A) \cdot P(B|A)$$

$P(B|A) = P(B)$ if A and B are independence

## 2. Bayes Theorem

A : sick
B : diagnosed sick

$P(A|B) = ?$

From Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(A) \cdot P(B|A)}{P(A \cap B) + P(A' \cap B)}$$

$$= \frac{P(A) \cdot P(B|A)}{P(A) \cdot P(B|A) + P(A') \cdot P(B|A')}$$

$$P(spam|lottery) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(A) \cdot P(B|A)}{P(A \cap B) + P(A' \cap B)} = \frac{1/5 \cdot 7/10}{1/5 \cdot 7/10 + 4/5 \cdot 1/8}$$

Prior $P(A)$

Event $E$

Posterior $P(A|E)$

Example:

Prior $P(spam) = \dfrac{P(spam)}{P(spam) + P(not\ spam)}$

Event Email contains "lottery"

Posterior $P(spam | lottery)$

Note: if $> 1$ event, Naive Bayes model may be consider

(Events are independent) - Assumption

## 3. Binomial distribution

$$X \sim Binomial(n, p)$$

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

$$\binom{n}{x} = \frac{n!}{(n-x)!\,x!}$$

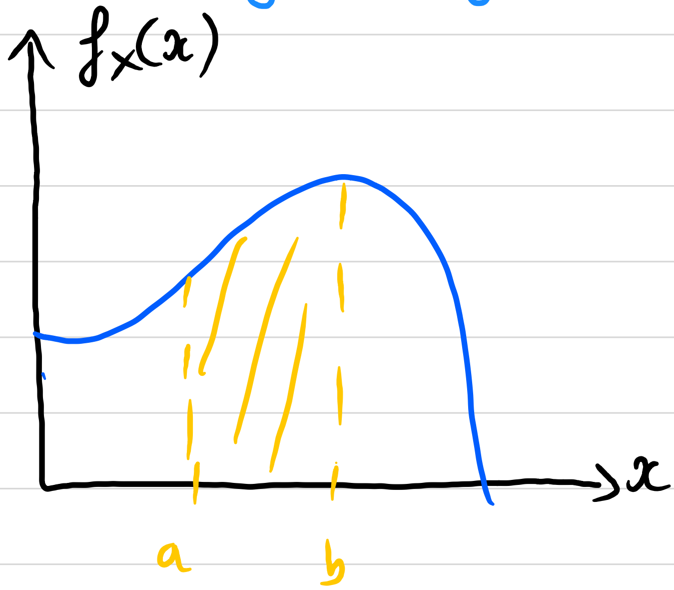
## 4. Bernoulli distribution

$$P(X=x) = \begin{cases} p & \text{if } x=1 \\ 1-p & \text{if } x=0 \end{cases}$$

Failure / Success

Probability Mass Function (PMF)

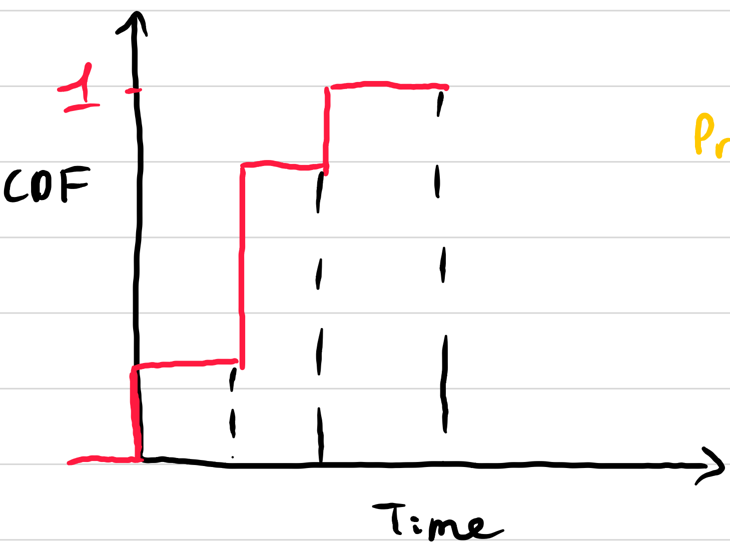
## 5. Probability Density Functions (PDF)




$f_X(x)$

$P(a < X < b) =$ area under $f_X(x)$

## 6. Cumulative distribution function (CDF)




$CDF(x) = P(X \leq x)$

Properties

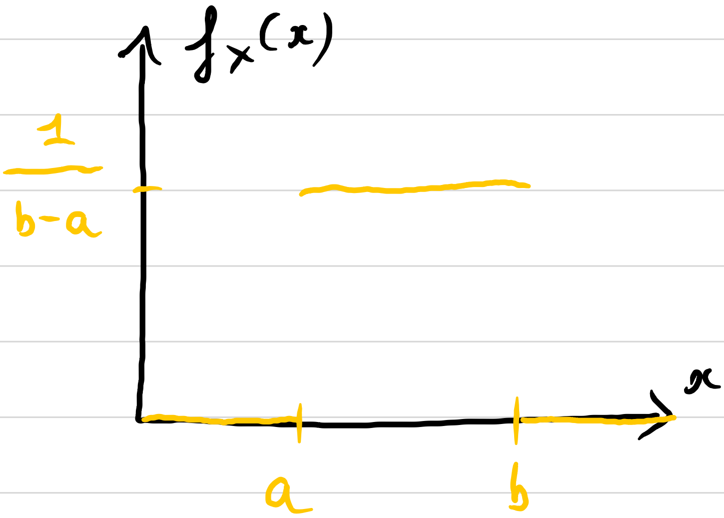
$0 \leq F_X(x) \leq 1$

Left endpoint is 0

Right endpoint is 1

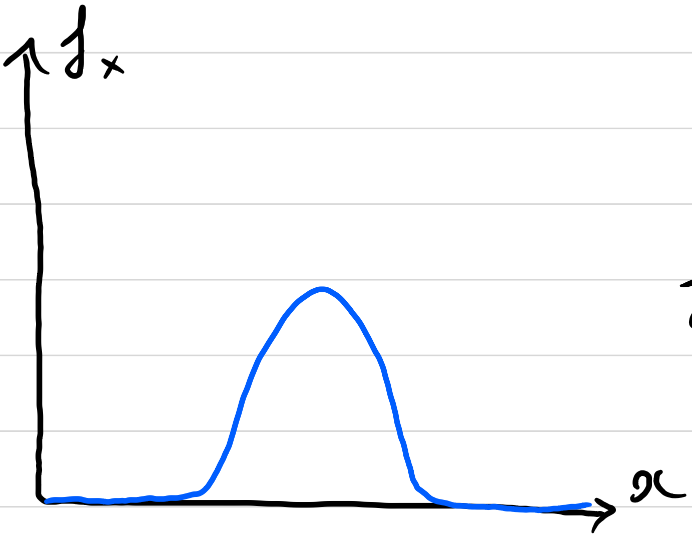
Never decreases

## 7. Uniform distribution




$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \notin (a, b) \end{cases}$$

## 8. Normal distribution




$$X \sim N(\mu, \sigma^2)$$

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

### Standardization

$$Z = \frac{X - \mu}{\sigma}$$

## 9. Chi-squared distribution

$$\begin{aligned} F_W(w) &= P(W \leq w) \\ &= P(Z^2 \leq w) \\ &= P(|Z| \leq w) \\ &= P(-\sqrt{w} \leq Z \leq \sqrt{w}) \end{aligned}$$

n dfs

$$W_K = \sum_{i=1}^{K} Z_i^2$$

## 10. Expected value

**Discrete random variables**

$$E[X] = \sum_{x} x \, p_X(x)$$

(PMF)

**Continuous random variables**

$$\int_{-\infty}^{\infty} x . f_X(x) \, dx$$

(PDF)

## 11. Variance

$$\sigma^2 = E\left[(X - E[X])^2\right]$$

$$= E[X^2] - E[X]^2$$

## 12. Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

## 13. Sum of Gaussians

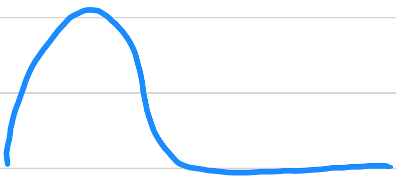
$$W = aX + bY$$

$$\begin{cases} X \sim N(\mu_X, \sigma_X^2) \\ Y \sim N(\mu_Y, \sigma_Y^2) \end{cases} \text{Independent} \Rightarrow W \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

## 14. Standardizing

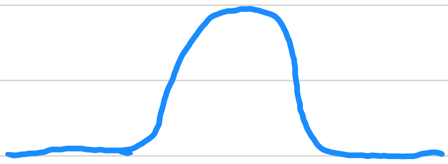
| | | |
|---|---|---|
| $X$ | $X-\mu$ | $\frac{X-\mu}{\sigma}$ |
| mean $=\mu$ | mean $=0$ | mean $=0$ |
| std $=\sigma$ | std $=\sigma$ | std $=1$ |

## 15. Skewness



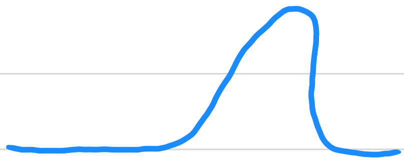
Positively skewed

$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]>0$



Not skewed

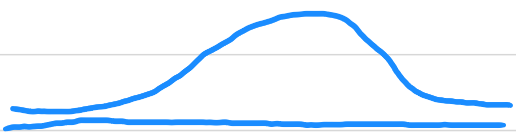
$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]=0$



Negatively skewed

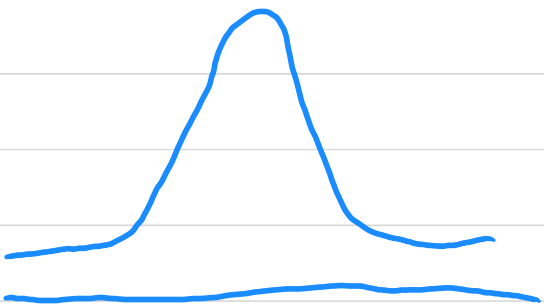
$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]<0$

## 16. Kurtosis (high and low)



Thin tails

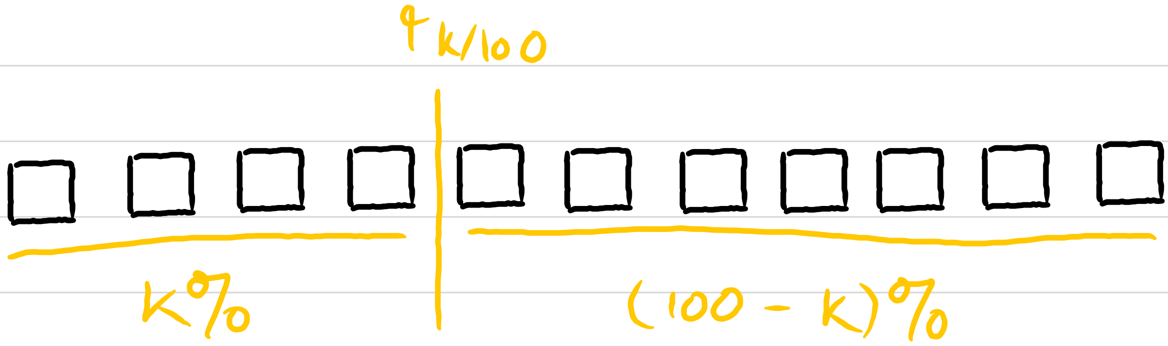
$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]=$ small



Thick tails

$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]=$ large

## 17. Quantiles




$k\%$ quantile ($q_{k/100}$) is the value such that

$$P(X \leq q_{k/100}) = \frac{k}{100}$$

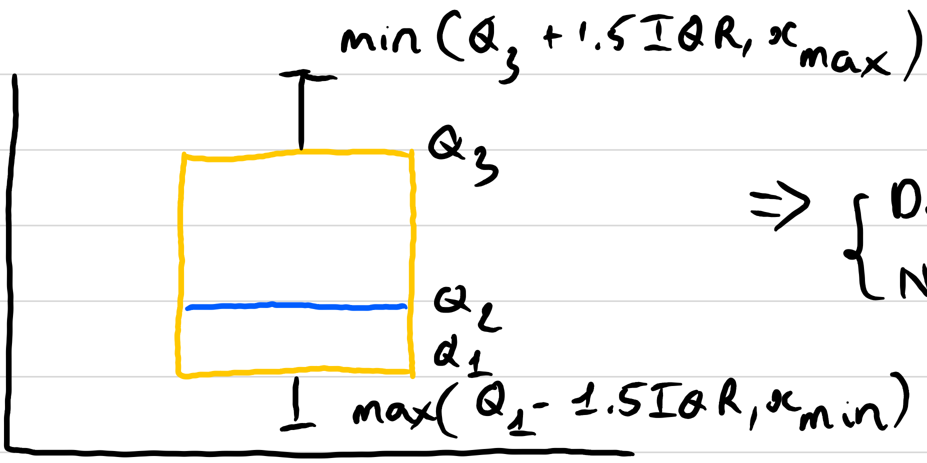
## 18. Box-plot

$Q_1 = q_{0.25}$ $\qquad x_{min}, x_{max}$

$Q_2 = q_{0.5}$

$Q_3 = q_{0.75}$

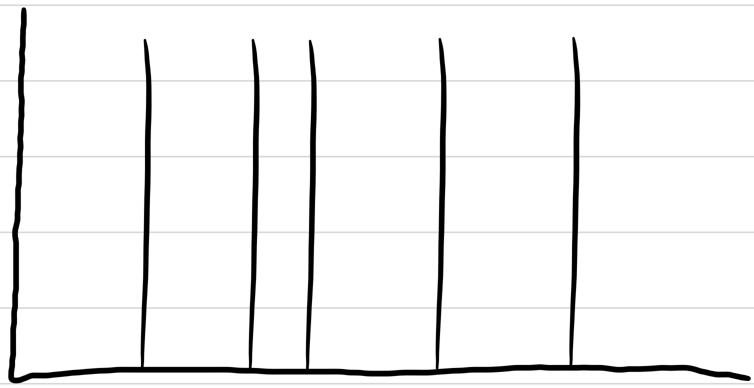
$IQR = Q_3 - Q_1 = q_{0.75} - q_{0.25}$



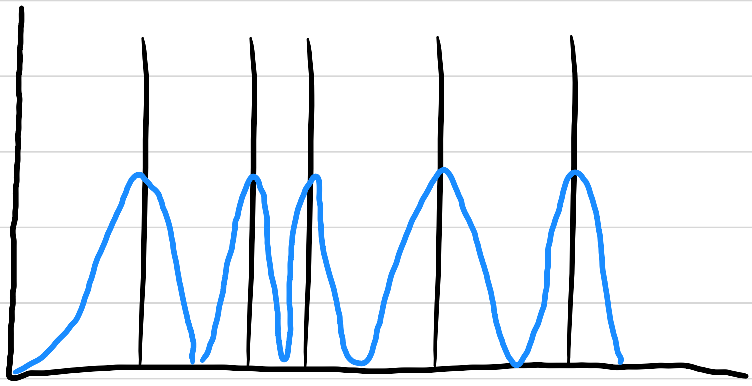

$\Rightarrow \begin{cases} \text{Data is skewed} \\ \text{No outliers} \end{cases}$

# 19. Kernel Density Estimation (KDE)

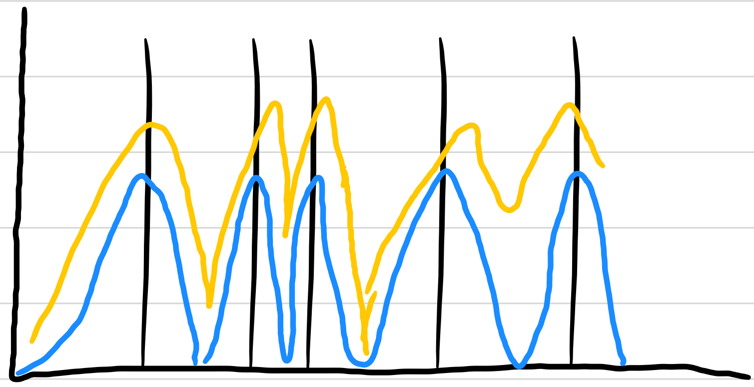
First: draw observations along the x axis



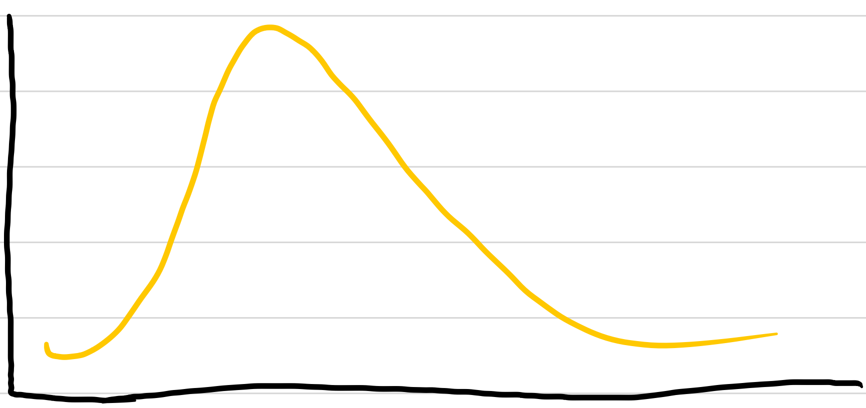
Second: draw a gaussian centered at each observation (can choose different σ)



Third: multiply everything by 1/n and sum the curves



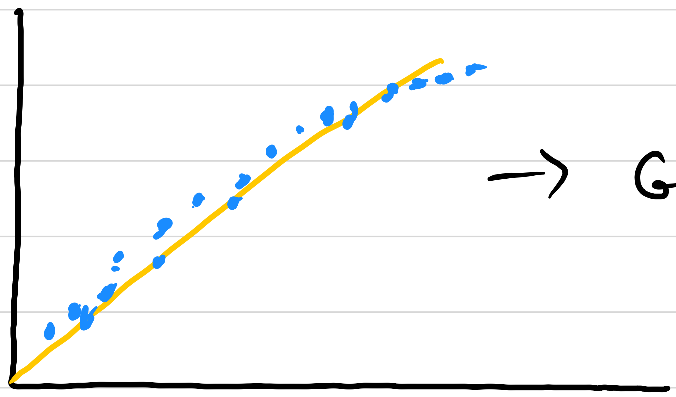
if use all datasets:



Approximated PDF based on the data

Violin plot = info of KDE + info of box plot

## 20. QQ plot (quantile-quantile plot)



→ Gaussian

## 21. Joint distribution

**Discrete**

$$P_{XY}(x,y) = P(X=x, Y=y) = \underbrace{P(x) \cdot P(y)}_{\text{independent}}$$

**Continuous** (scatter)

## 22. Marginal distribution

- Distribution of 1 variable while ignoring the other

Ex: age and height, only care about joint distribution of age

$$P_Y(y_j) = \sum_i P_{XY}(x_i, y_i)$$

## 23. Conditional distribution

Joint PDF

**Discrete:**

$$p_{Y|X=x}(y) = \frac{p_{XY}(x, y)}{p_X(x)}$$

Marginal distribution

**Continuous:**

$$f_{Y|X=x}(y) = \frac{f_{XY}(x, y)}{f_X(x)}$$

## 24. Covariance

**equal probs**

$$\operatorname{Cov}(X, Y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n}$$

**unequal probs**

$$\operatorname{Cov}(X, Y) = \sum p_{XY}(x_i; y_i)(x_i - \mu_x)(y_i - \mu_y)$$

$$\operatorname{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

**Covariance matrix**

| | A | B |
|---|---|---|
| A | Var(A) | Cov(A, B) |
| B | Cov(B, A) | Var(B) |

## 25. Correlation coefficient

$$Corr = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

$$\frac{15}{100} : \frac{3}{10} = \frac{15}{30}$$

## 26. Law of large numbers

$n$ : number of samples

$X_i$ : is the i-th random sample from the population

Each $X_i$ are independent and identically distributed

as $n \to \infty$ $\quad \frac{1}{n}\sum_{i=1}^{n} X_i \to E[X] = \mu_X$

## 27. Central Limit Theorem

<u>Discrete random variable</u>

as $n$ become sufficiently large we will get a normal distribution with parameters

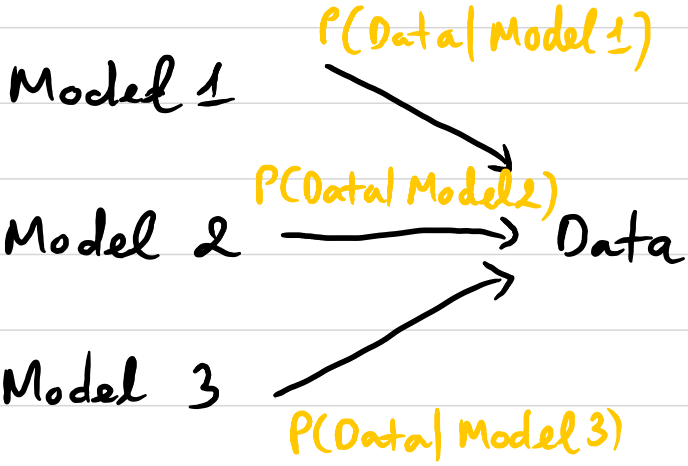
$$\mu = np$$
$$\sigma^2 = np(1-p)$$

Continuous random variable

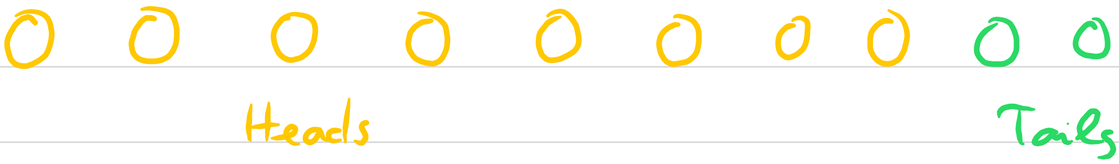
$$\text{As } n \to \infty \qquad \frac{\sum_{i=1}^{n} X_i - n\,E[X]}{\sigma_X \sqrt{n}} \sim N(0, 1^2)$$

## 28. Maximum Likelihood Estimation

- It means we pick the scenario that made the evidence more likely




Maximize
P(Data| Model)




$p = P(H)$

→ Likelihood $L(p; 8H) = p^8 (1-p)^2$

Want $p$ that maximizes the chances of seeing 8H

→ Log-likelihood $\ell(p; 8H) = \log\left(p^8 (1-p)^2\right)$

$$= 8 \log p + 2 \log(1 - p)$$

$$\frac{d}{dp} l(p, 8H) = \frac{8}{p} - \frac{2}{1-p} = 0$$

$$\Rightarrow \hat{p} = \frac{8}{10}$$

General cases:

n coins $\quad X = (X_1, \dots, X_n)$

k heads

$$X_i \overset{iid}{\sim} \text{Bernoulli}(p)$$

$$l(p; x) = \left(\sum_{i=1}^{n} x_i\right) \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p)$$

To find the maximum:

Find p for $\frac{d}{dp} l(p; x) = 0$

where,

$$\frac{d}{dp} l(p; x) = \frac{\sum_{i=1}^{n} x_i}{p} + \frac{n - \sum_{i=1}^{n} x_i}{1-p}(-1) = 0$$

$$\Rightarrow p = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

## 29. Regularization

Model: $y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$

Log-loss: $ll$

L2 Regularization error: $a_n^2 + a_{n-1}^2 + \dots + a_1^2$

Regularization parameter: $\lambda$

Regularized error: $ll + \lambda(a_n^2 + a_{n-1}^2 + \dots + a_1^2)$

## 30. Frequentist vs. Bayesian statistics

| Frequentists | Bayesians |
| --- | --- |
| - Probabilities represent long term frequency of events | - Probabilities represent the degree of belief (or certainty) |
| - Concept of Likelihood | - Concept of Prior |
| - Goal: Find the model that most likely generated the observed data | - Goal: update prior belief based on observations |

## 30. Maximum a Posteriori (MAP)

1 value for the parameter -> Choose the one with highest probability -> Mode of the $\boxed{\text{updated belief}}$

Posterior

## 31. Bayesian Statistics

Coin flip case

| "Fair" | "Biased" | "Mystery" |
|---|---|---|
| $P(H) = 0.5$ | $P(H) = 0.8$ | Either fair of biased |

Event to predict | Evidence

A -> Y takes some value | B -> X take some value

Y : odds of H for your coin | X : result of coin flip

$$Y = \begin{cases} 0.5 & \text{if coin is fair} \\ 0.8 & \text{if coin is biased} \end{cases} \qquad X = \begin{cases} 0 & \text{if } T \\ 1 & \text{if } H \end{cases}$$

$x = \underline{1}$

$$\boxed{\begin{aligned} P(Y = 0.5) &= 0.75 \\ P(Y = 0.8) &= 0.25 \end{aligned}}$$ -> Priors

Posterior

$$P(Y=0.5 \mid X=1) = \frac{P(X=1 \mid Y=0.5) \cdot P(Y=0.5)}{P(X=1 \mid Y=0.5) \cdot P(Y=0.5) + P(X=1 \mid Y=0.8) \cdot P(Y=0.8)}$$

$$= 0.652 \Rightarrow P(Y=0.8 \mid X=1) = 0.348$$

Key summary :

- Bayesians update prior beliefs

- MAP with uninformative priors is Just MLE

- With enough data, MLE and MAP estimates usually converge

- Good for instances when you have limited data or strong prior beliefs

- Wrong priors, wrong conclusions

## 32. Margin of error

$$\frac{x-\mu}{\sigma} = Z \sim N(0, 1^2)$$

standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Margin of error $z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Confidence interval : $\bar{x} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

## 33. Confidence interval for proportions

$$\hat{p} = \frac{x}{n}$$

CI : $\hat{p} \pm$ margin of error

margin of error : $z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

## 34. Hypothesis testing

| | Reality | |
|---|---|---|
| Decision | $H_0$ true | $H_0$ false |
| Reject $H_0$ | Type I error | Correct |
| Don't reject $H_0$ | Correct | Type II error |

- Type I error is more severe

Tct 171103

**Decision rule**

if p-value $< \alpha$ : reject $H_0$

if p-value $> \alpha$ : do not reject $H_0$