Probabilities Recap

10-605 Machine Learning with Large Datasets

Fall 2022

Outline

- Setup
- Random variables
- Distribution function
- Expectation
- Multivariate Distributions
- Independence
- ROC curve
- Probability in Hashing (birthday paradox)

Setup

- **Sample Space**
  - A set of all possible outcomes or realizations of some random trial.
- **Event**
  - A subset of sample space
- **Probability Axioms**
  - $P(A) \geq 0$ for every A
  - $P(\Omega)=1$;
  - If A1, A2, . . . are disjoint, then

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Random variables

- **Definition**
  - A random variable is a function that maps from the sample space to the reals ($X : \Omega \rightarrow R$), i.e., it assigns a real number $X(\omega)$ to each outcome ω.
- **Example**
  - X returns 1 if a coin is heads and 0 if a coin is tails. Y returns the number of heads after 3 flips of a fair coin.
- Random variables can take on many values, and we are often interested in the distribution over the values of a random variable, e.g., $P(Y = 0)$

Distribution function

- **Definition**
  - Suppose X is a random variable, x is a specific value that it can take,
  - Cumulative distribution function (CDF) is the function $F : R \rightarrow [0, 1]$, where $F(x) = P(X \leq x)$.

- **If X is discrete ⇒ probability mass function: f(x) = P(X = x).**

Distribution function (cont.)

- If X is continuous ⇒ probability density function for X if there exists a function f such that f(x) ≥ 0 for all x,

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

and for every a ≤ b,

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$$

If F(x) is differentiable everywhere, f(x) = F'(x).

Example of distributions

Discrete variable	Probability function	Mean	Variance
Uniform $X \sim U[1, \ldots, N]$	$1/N$	$\frac{N+1}{2}$	
Binomial $X \sim Bin(n, p)$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	np	
Geometric $X \sim Geom(p)$	$(1-p)^{x-1} p$	$1/p$	
Poisson $X \sim Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^x}{x!}$	λ	
Continuous variable	Probability density function	Mean	Variance
Uniform $X \sim U(a, b)$	1/ (b-a)	(a + b)/2	
Gaussian $X \sim N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$	μ	
Gamma $X \sim \Gamma(\alpha, \beta)$ $(x \geq 0)$	$\frac{1}{\Gamma(\alpha)\beta^a} x^{a-1} e^{-x/\beta}$	$\alpha\beta$	
Exponential $X \sim exponen(\beta)$	$\frac{1}{\beta} e^{-\frac{x}{\beta}}$	β	

Expectation

- **Expected Values**
  - Discrete random variable X

$$E[g(X)] = \sum_{x \in \chi} g(x) f(x)$$

  - Continuous random variable X

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Expectation (cont.)

- **Mean and variance**

$$\mu = E(X)$$

$$var[X] = E[(X - \mu)^2]$$

We also have

$$var[X] = E[X^2] - \mu^2$$

Multivariate Distributions

- **Definition**

$$F_{X,Y}(x,y) := P(X \leq x, Y \leq y)$$

and

$$f_{X,Y}(x,y) := \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

- **Marginal Distribution of X (discrete case)**

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x,y)$$

What about continuous variable?

Independence

- **Independent Variables**
  - X and Y are independent if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Or

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Independence (cont.)

- **IID variable**
  - Independent and identically distributed (IID) random variables are drawn from the same distribution and are all mutually independent.

- **Linearity of Expectation**
  - Even if the events are not independent, this property still holds

$$E[\sum_{x=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$$

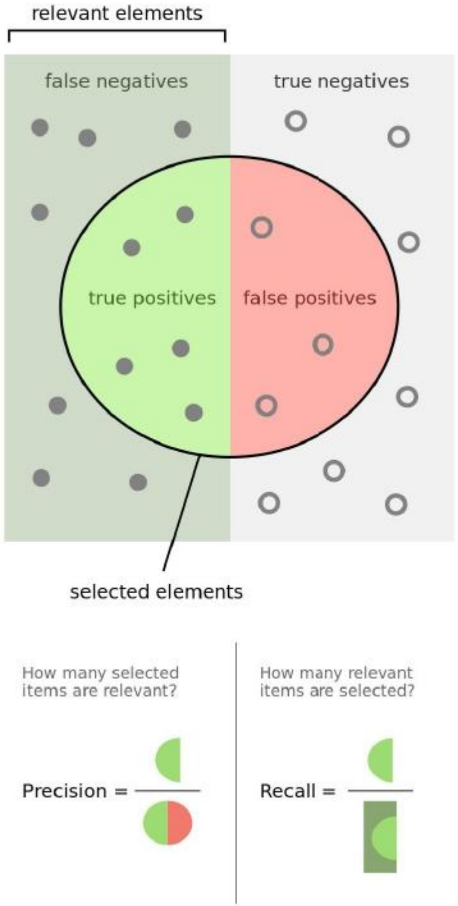
ROC curve

- **Confusion matrix**

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

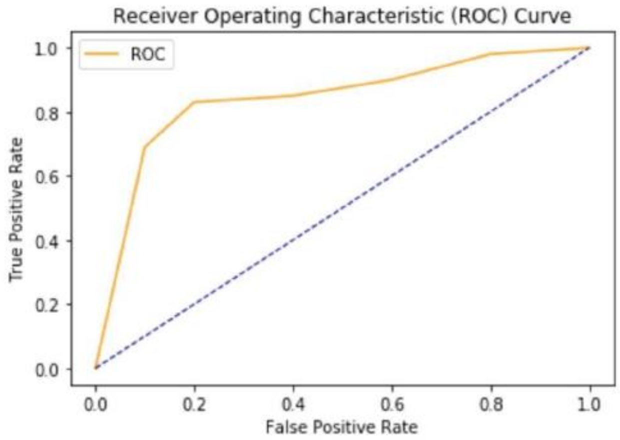
ROC curve

- **Statistics Computed from Confusion Matrix**
    - Precision: Out of all the predicted positive instances, how many were predicted correctly.
    - Recall: Out of all the positive classes how many instances were identified correctly.

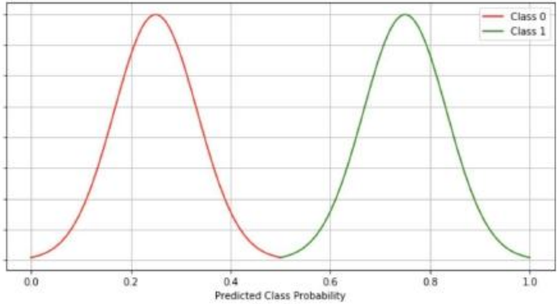


ROC curve

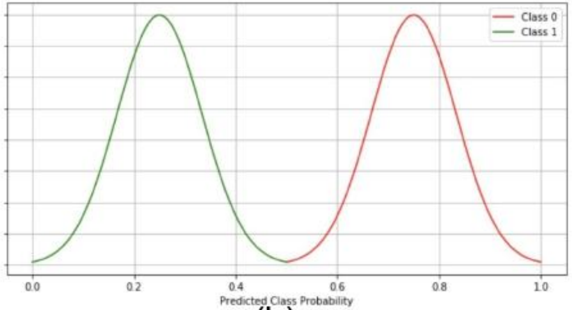
- **Introduction to AUC - ROC Curve**
  - how good the model is for distinguishing the given classes, in terms of the predicted probability



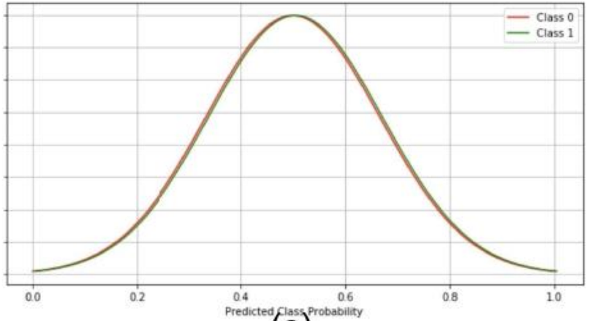
ROC curve



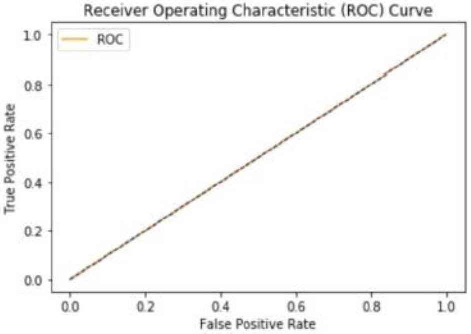
(a)



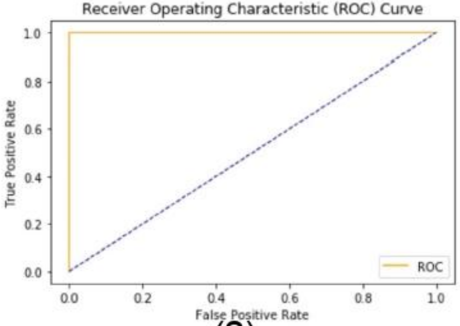
(b)



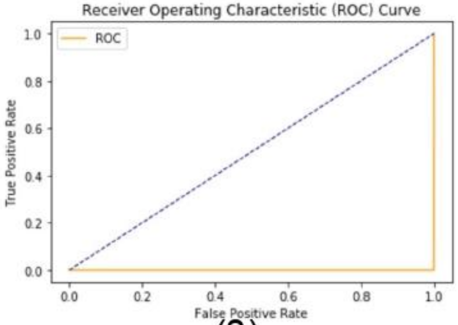
(c)



(1)



(2)



(3)

Probability in Hashing

Probability in Hashing

- **Assumption**
  - n=number of people
  - k=365
  - P(person *i* is born on day *j*) = 1/k

We are interested in the event A that **at least two people have the same birthday**.

$$
\begin{aligned}
P(A) &= 1 - P(\bar{A}) \\
&= 1 - \frac{k}{k} \cdot \frac{k-1}{k} \cdot \ldots \cdot \frac{k-n+1}{k} \\
&= 1 - \frac{k!}{(k-n)!k^n}.
\end{aligned}
$$

Probability in Hashing

- **Hashing**
  - Similar to assignments of birthdays
  - n items mapped into k slots
- **Hashing problems dealing with probabilities**
  - the expected number of items mapping to same slot
  - the expected number of empty slots
  - the expected number of collisions

Probability in Hashing

- **Empty slots**
  - The probability that slot j remains empty after mapping all n items is

$$(1 - \frac{1}{k})^n$$

  - The expected number of empty slots is

$$E(X) \;=\; \sum_{j=1}^{k} E(X_j) \;=\; k\left(1 - \frac{1}{k}\right)^n .$$

  - If k = n, we can get a max limitation of 0.367

KL Divergence

Question:
How different are two probability distributions from each other?

$$D_{KL}(P||Q) = \mathbb{E}_{x \sim P(\cdot)}[\log(\frac{P(X)}{Q(X)})] = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

It captures what is the expected “excess surprise” from using Q as a model for data when the actual distribution is P.

KL-divergence is NOT symmetric!

Concentration Inequalities

Can we figure out the probability that a random variable deviates from it's mean by a particular value; i.e. with how much probability does the following statement occur:

$$|\bar{X} - \mathbb{E}[X]| \leq \delta$$

Concentration inequalities are a family of such statements that provide **exact** bounds on this probability.

Some common ones are; Markov's, Chebyshev's, Hoeffding's, Chernoff's Bounds etc.

Markov’s Inequality

If X is non-negative, then for a positive value of a;

If $X \geq 0$,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Chebyshev’s Inequality

For a random variable X, with finite mean, and non-zero variance;

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Qualitatively, this statement tells us the probability that the value of a random variable deviates from it’s mean by ‘k’ standard deviations is bounded by 1/k^2.

Johnson and Lindenstrauss Lemma

Lemma For any $0 < \epsilon < 1$ and any interger n let k be a possitive interger such that

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n \tag{2}$$

then for any set A of n points $\in \Re^d$ there exists a map $f : \Re^d \rightarrow \Re^k$ such that for all $x_i, x_j \in A$

$$(1-\epsilon)||x_i - x_j||^2 \leq ||f(x_i) - f(x_j)||^2 \leq (1+\epsilon)||x_i - x_j||^2 \tag{3}$$

Note: The proof involves Markov's inequality