Department of Statistics, Seoul National University

Machine Learning and Statistical Inference

SCSC 2022 ML/DL SIG

October 11, 2022

Kyeongwon Lee

Outline

Artificial intelligence

Figure 1: Artificial intelligence in imagination and reality.

Trends

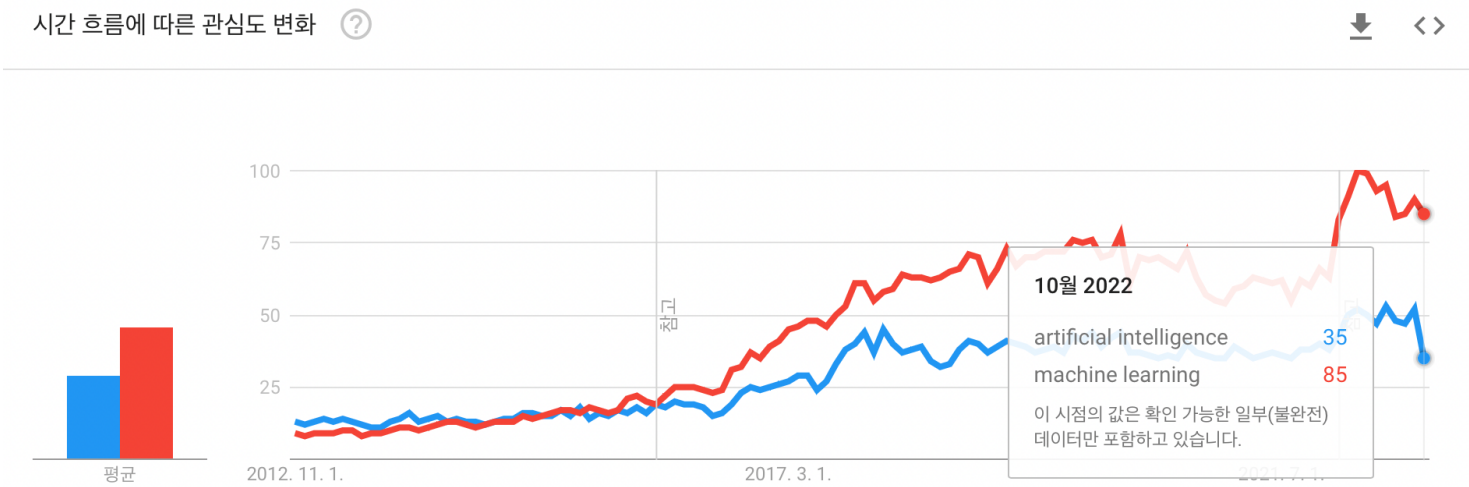


Figure 2: Trends in search frequency of AI and ML over the past decade.

History

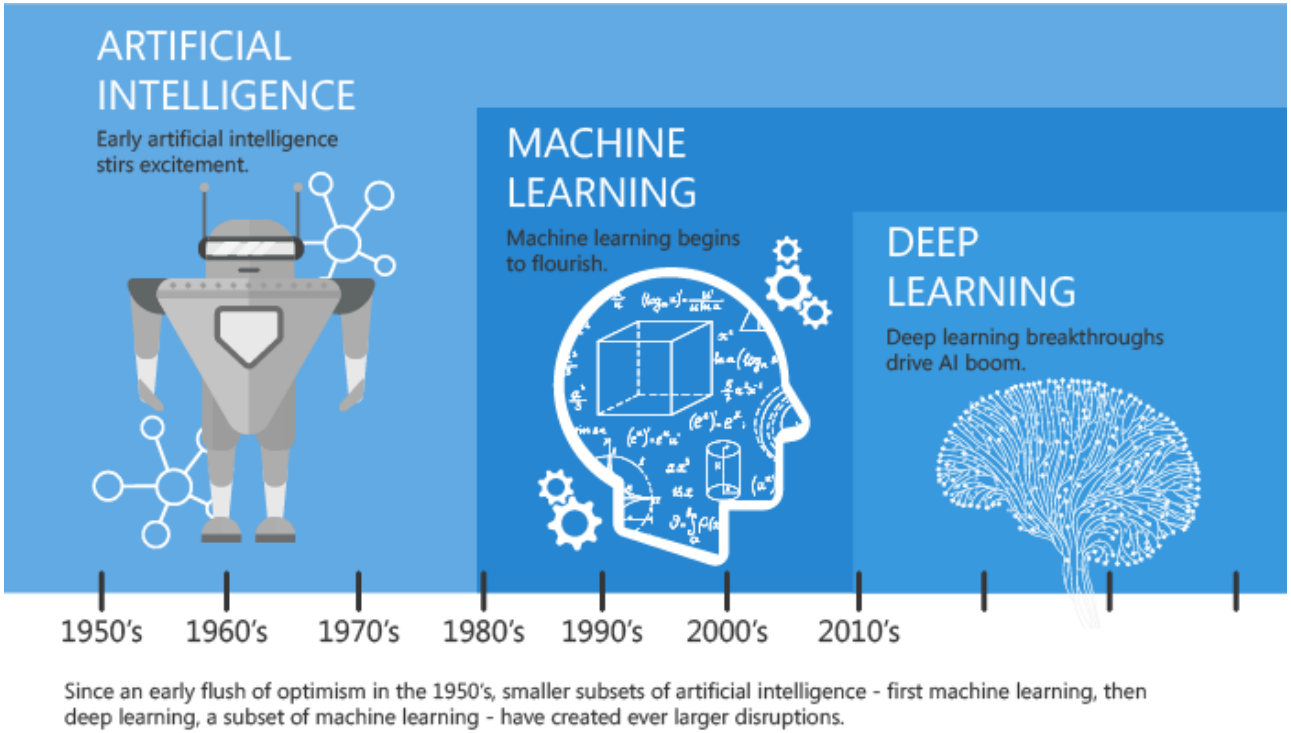
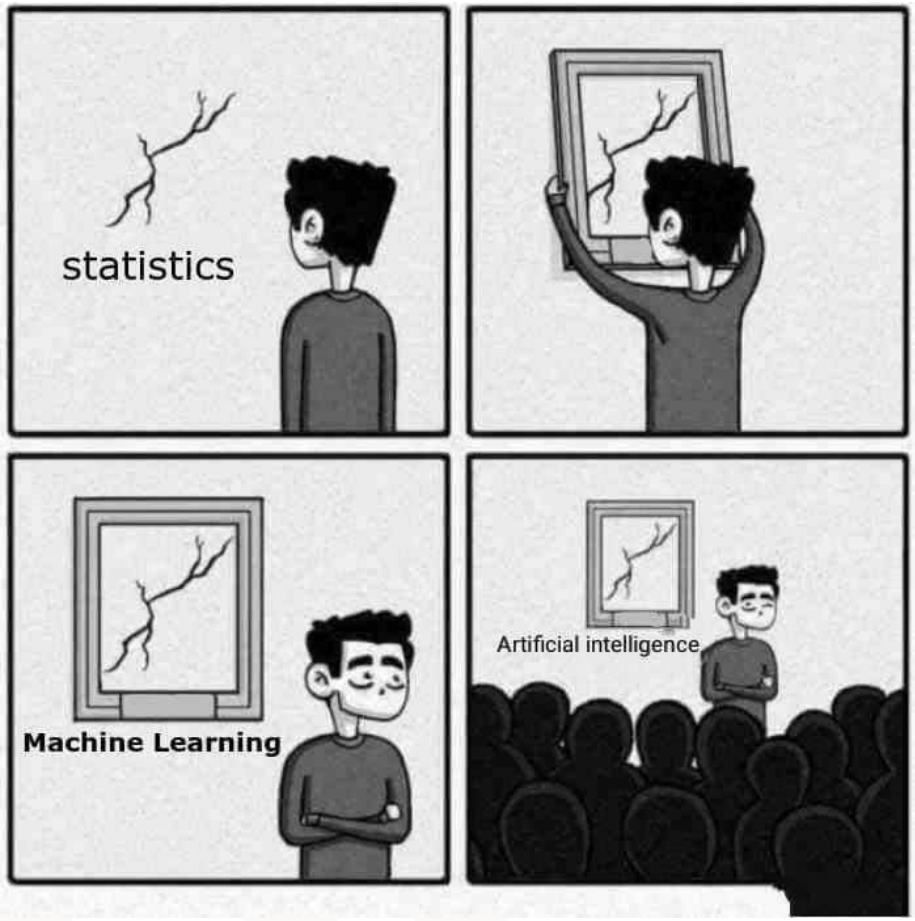


Figure 3: History from artificial intelligence to deep learning.

Statistics



Outline

Machine learning

- Machine learning (ML) is a field of inquiry devoted to understanding and building methods that 'learn'.
- that is, methods that leverage data to improve performance on some set of tasks.

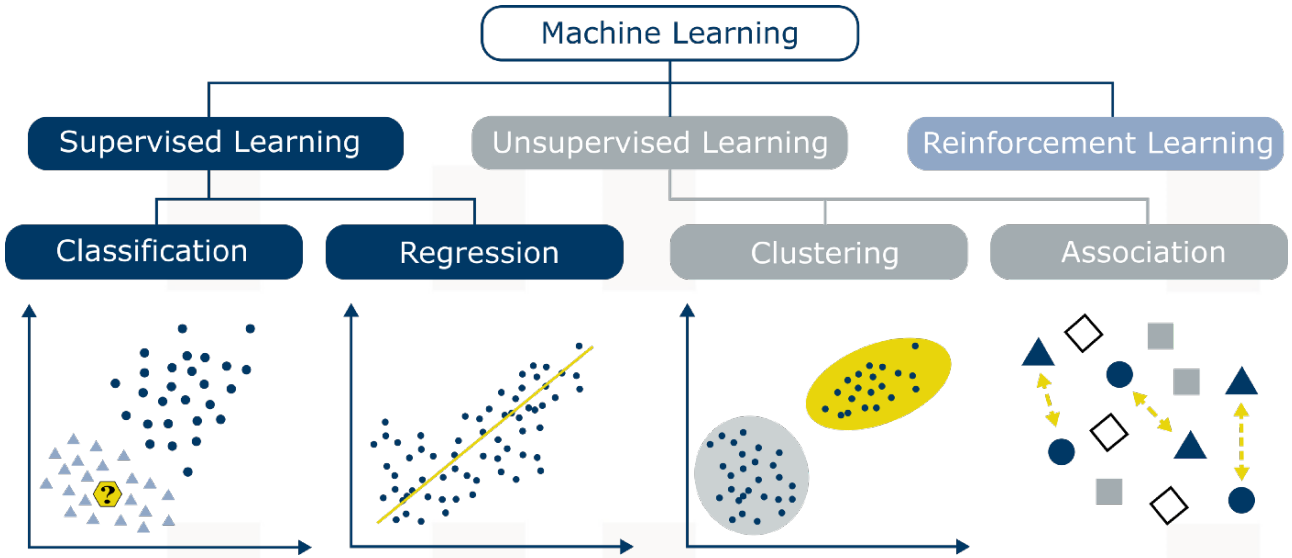


Figure 4: Types of machine learning. Source: TDS.

Supervised learning

Problems

- We have training data $\{(x_1, y_1), \cdots, (x_n, y_n)\}$.
- We want to find the relationship between the input (x) and output (y).
- We want to predict the output y^{new} from new input x^{new}.

Regression vs Classification

Regression The output y is numeric.

Classification The output y is categorical.

Example - binary classification

- x_i's are image.
- $y_i \in \{\text{“}cat\text{”}, \text{“}dog\text{”}\}$ (discrete).

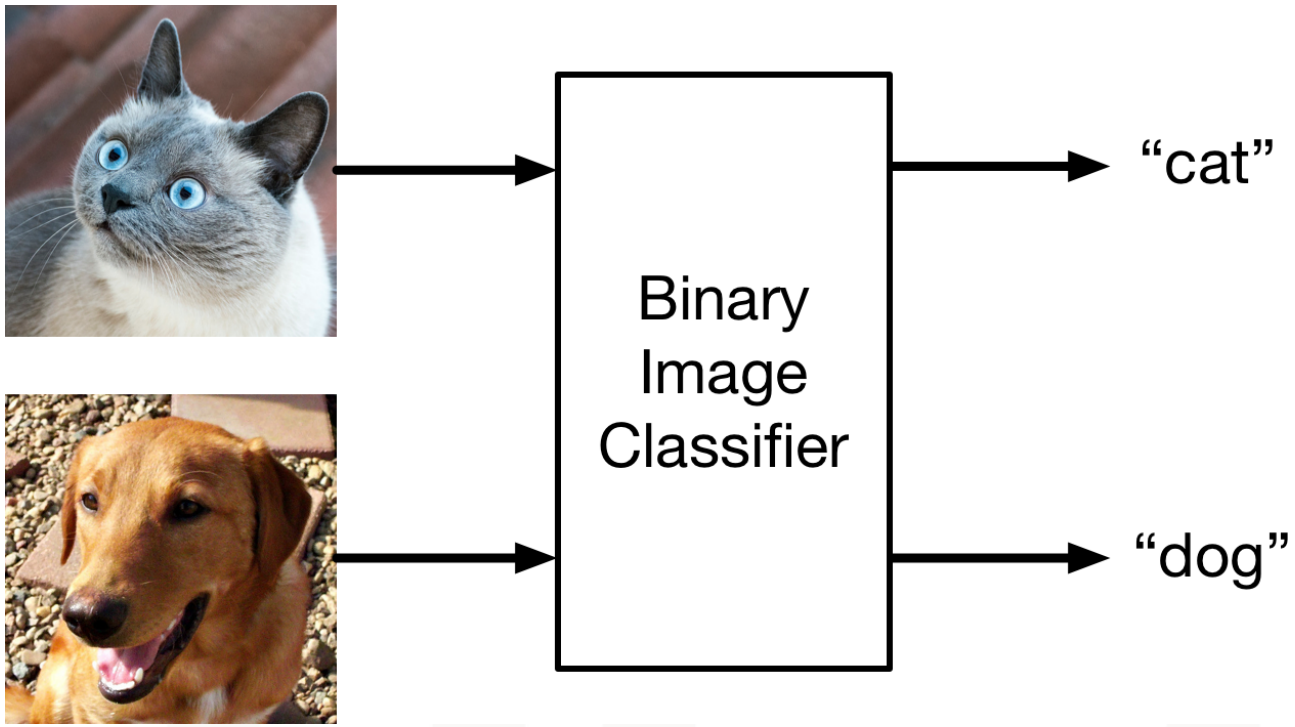


Figure 5: A binary classifier for cats and dogs

Procedure

The supervised learning process consists of these steps.

Modeling First, we should determine the model for problem.

Evaluation Next, we need to point the machine in the direction it should learn. (good performace $\rightarrow$ higher score / lower loss)

Fitting Fit the model to get a high score or low loss.

Prediction Predict the output y^{new} from new input x^{new} and the fitted model.

Example - binary classification

Modeling Consider a convolutional neural networks model f such as ResNet (He, Zhang, Ren, & Sun, 2016) or VGGNet (Simonyan & Zisserman, 2014).

Evaluation Consider the following binary cross entropy (BCE) loss

$$\text{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(f(x_i)) + (1-y_i)\log(1-f(x_i))\right] \quad (1)$$

Fitting Find the model which has lowest BCE loss using optimization algorithm such as Adam (Kingma & Ba, 2014).

Prediction Predict whether the output y^{new} is cat or not from new input image x^{new} and the fitted model.

Outline

Descriptive statistics

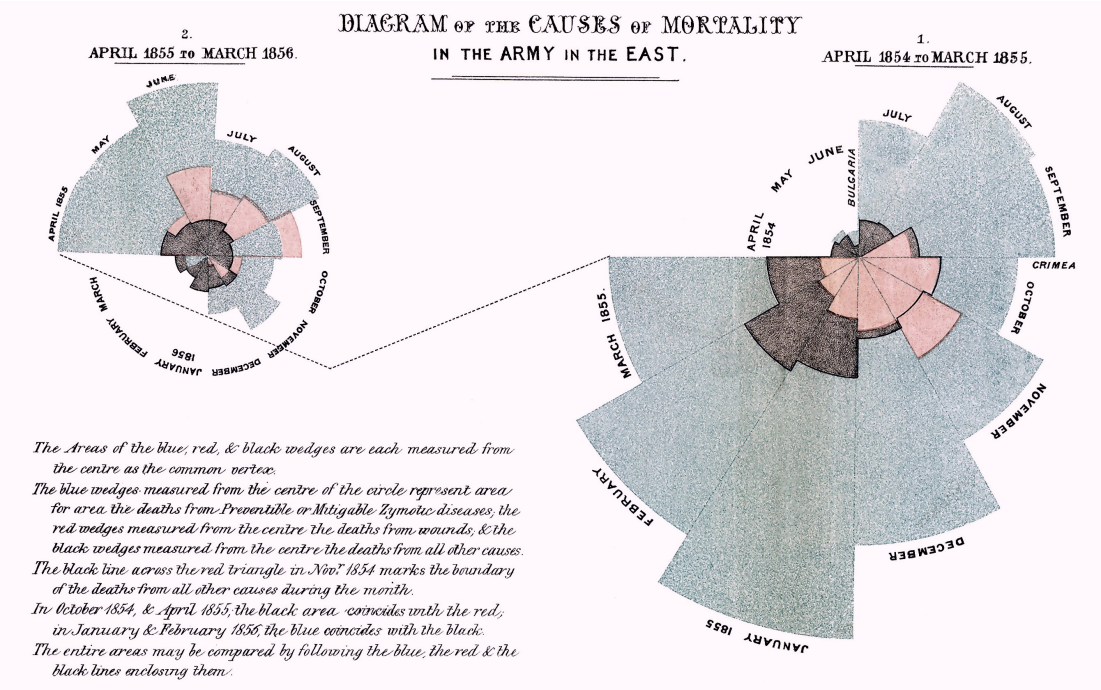


Figure 6: Descriptive statistics is the process of using and analyzing statistics which quantitatively describes or summarizes features from a collection of information.

Inferential statistics

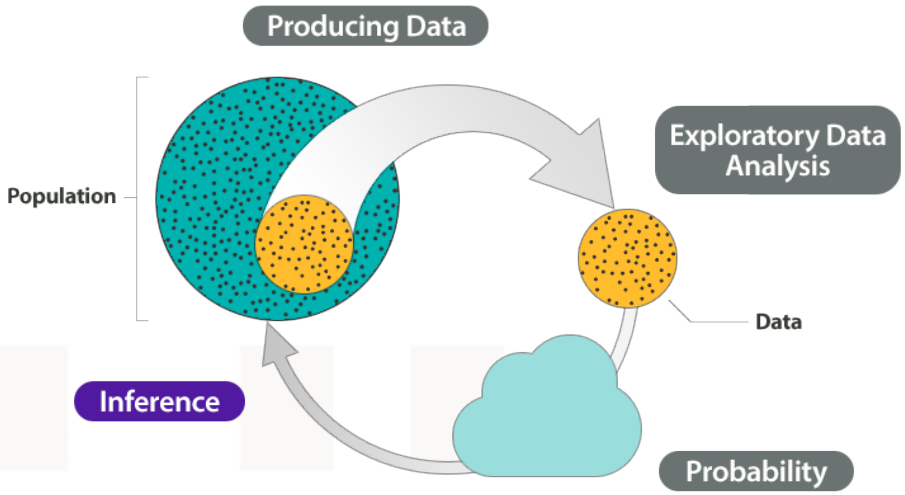


Figure 7: Inferential statistics is the process of using data analysis to infer properties of an underlying probability distribution of the population from the sample.

Statistical model

- A statistical model $(\mathcal{S}, \mathcal{P})$ is a mathematical model $\mathcal{P}$ that embodies a set of statistical assumptions concerning the generation of sample data $\mathcal{S}$.
- A statistical model represents, often in considerably idealized form, the data-generating process.

Example 1 (Normal model)

$$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \quad \mu \in \mathbb{R},\ \sigma^2 > 0, \tag{2}$$

where $i.i.d.$ means independently and identically distributed.

- In the above example, the parameters μ and σ determine (and represent) probability distribution of the model.
- In statistics, a parameter of the model is a variable which represent statistical model (or population).

Example - Binary classification

For some $f \in \mathcal{F}$, assume a model

$$y|x \sim Ber(f(x)). \tag{3}$$

Here, $Ber(p)$ is the Bernoulli distribution which has the probability mass function

$$Z \sim Ber(p) \Longleftrightarrow \mathbb{P}(Z = z) = p^z(1-p)^{1-z}, \quad z \in \{0, 1\}. \tag{4}$$

Note that the function f is a parameter of the model.

Parametric and nonparametric model

Suppose that we have a statistical model $(\mathcal{S}, \mathcal{P})$ with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

- The model is said to be parametric if Θ has a finite dimension.
- The model is said to be nonparametric if Θ has a infinite dimension.

Example 2 (Normal model)

Consider the model

$$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \quad \mu \in \mathbb{R},\ \sigma^2 > 0,$$

where $i.i.d.$ means independently and identically distributed. Since we can parameterize the model space as follows

$$P_\theta = \{P_\theta = N(\mu, \sigma^2) : \theta = (\mu, \sigma^2) \in \Theta\}, \quad \Theta = \mathbb{R} \times \mathbb{R}_{>0},$$

the model is parametric.

Example - Binary classification

For some $f \in \mathcal{F}$, assume a model

$$y|x \sim Ber(f(x)). \tag{5}$$

- (Linear) If $x \in \mathbb{R}^d$ and

$$\mathcal{F} = \{f(x) = a + x^T b : a \in \mathbb{R},\ b \in \mathbb{R}^d\},$$

the model is parametric.

- (Neural network) If

$$\mathcal{F} = \{f_\theta(x) : f_\theta \text{ is neural network with weight and bias parameters } \theta.\},$$

the model is parametric.

- If

$$\mathcal{F} = \{f(x) : f \text{ is continuous.}\},$$

the model is nonparametric

Comparison to Machine Learning

The supervised learning process consists of three main steps.

Modeling First, we should determine the model for problem.
Express the relationship between x and y as a statistical model.

Evaluation Next, we need to point the machine in the direction it should learn. (good performace $\rightarrow$ higher score / lower loss)
Decide **how to estimate** the parameter.

Fitting Fit the model to get a high score or low loss.
Estimate the parameter of the model from the sample.

Prediction Predict the output y^{new} from new input x^{new} and the fitted model.
Sample new y^{new} from new input data x^{new}.

Outline

Frequentist and Bayesian inference

- The frequentist and Bayesian inference are statistical inference based on the frequentest statistics and Bayesian statistics, respectively.
- The main difference between the two begins in how you interpret the probability.
- In short, frequentist assumes objective probability and Bayesian assumes subjective probability.

Example - classification

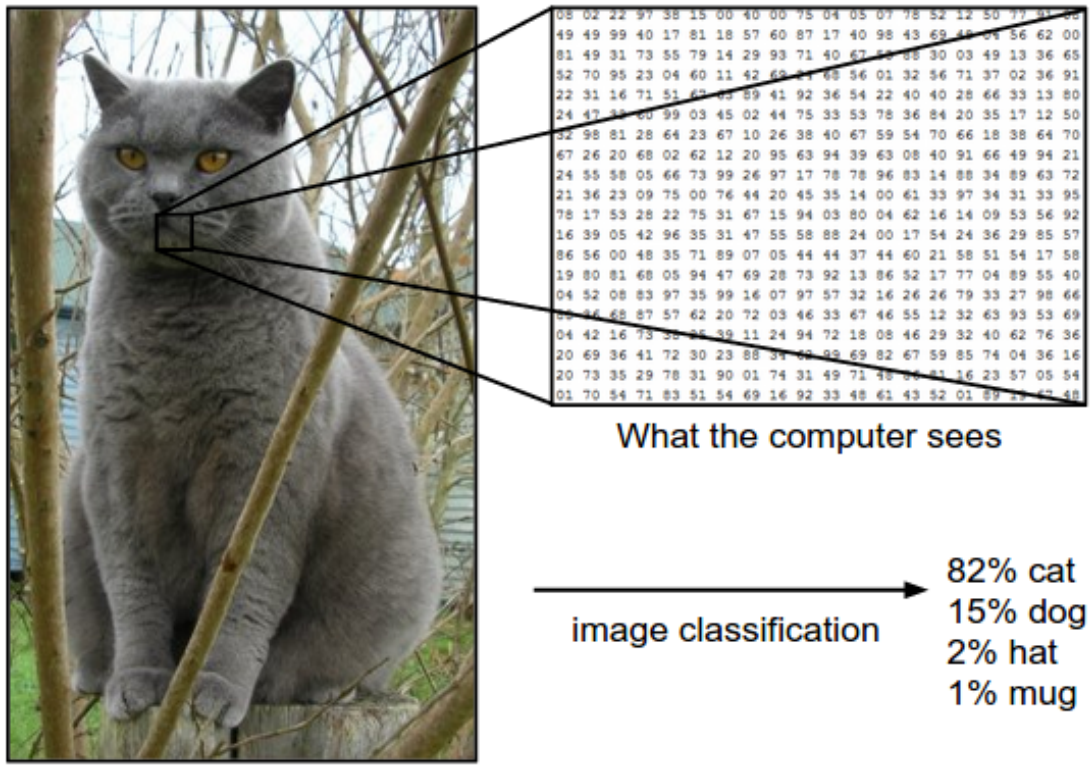


Figure 8: What does an 82% chance of being a cat mean? Source: CS231n.

Classical probability

"The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible"

- Pierre-Simon Laplace

Classical probability

- The classical probability assume that each event is expected to the same extent.
- When each event is expected to equally and cannot occur simultaneously, the classical probability of the event A is defined as

$$\mathbb{P}(A) = \frac{\text{the number of cases of } A\text{occurs}}{\text{the number of all possible cases}}$$

- For example, the probability of rolling a 1 when a fair die is rolled is $1/6$.

Frequentist statistics

- Frequentist statistics treats "probability" in equivalent terms to "frequency" and draws conclusions from sample-data by means of emphasizing the relative frequency or proportion of findings in the data.
- Frequentist assume the experiments can be repeat infinitely and define the probability as the limit of the relative frequency.
- The probability of rolling a 1 when a die is rolled is $1/6$, meaning that if the die is tossed infinitely, the relative frequency of rolling a 1 is about $1/6$.
- Probability is objective and inductive.
- Parameter is a fixed constant.

Frequentist statistics

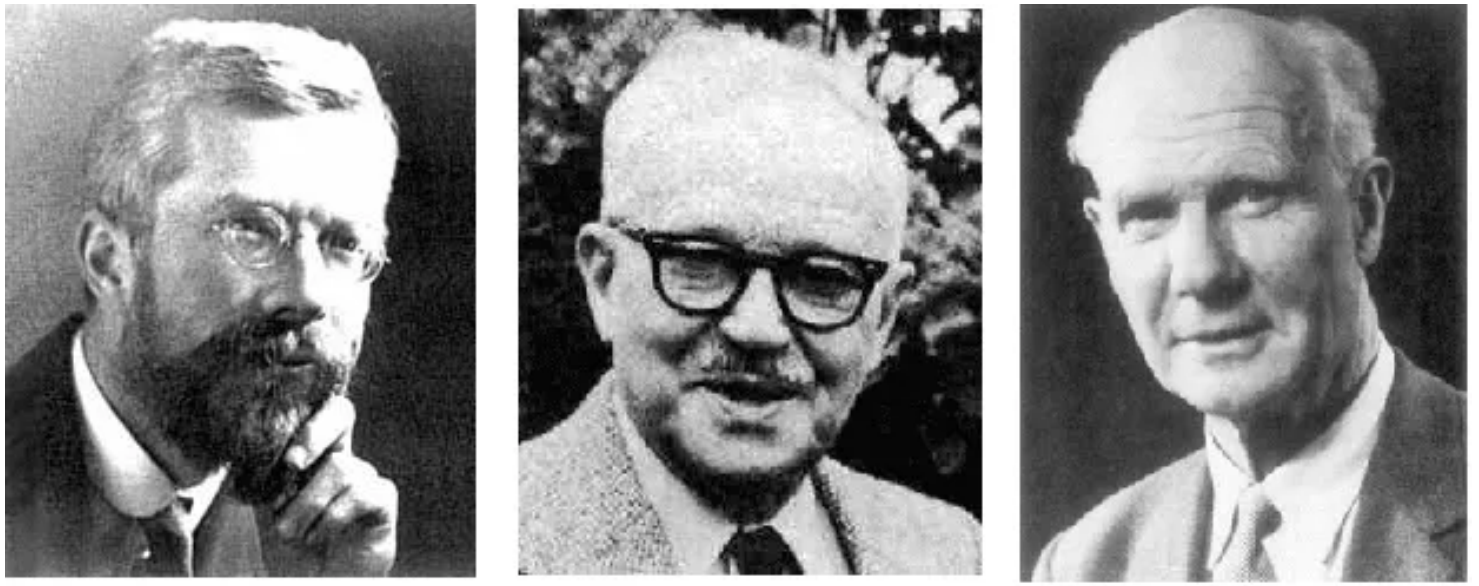


Figure 9: Ronald Fisher, Jersey Neyman and Egon Pearson

Maximum likelihood estimation

Assume a model has a density function $f(D_n; \theta)$ for the parametric model P_θ and a dataset D_n. Consider following likelihood function

$$L_n(\theta) = f(D_n; \theta), \tag{6}$$

which is a function of θ. A maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space Θ, that is

$$\hat{\theta} = \arg\max L_n(\theta). \tag{7}$$

Example - binary classification

For some $f_\theta,\ \theta \in \Theta$, assume a model

$$y_i | x_i \sim Ber(f_\theta(x_i)). \tag{8}$$

Then a maximum likelihood estimator of the model is given by

$$\hat{\theta} = \arg\max \sum_{i=1}^{n} \left[y_i \log f_\theta(x_i) + (1 - y_i) \log f_\theta(x_i) \right]. \tag{9}$$

That is, machine learning model that minimize BCE loss is a maximum likelihood estimation for the Bernoulli model.

Bayesian statistics

Figure 10: Thomas Bayes

Bayesian statistics

- Bayesian statistics treats "probability" as a degree of belief in an event.
- The degree of belief may be based on prior knowledge about the event, such as the results of previous experiments, or on personal beliefs about the event.
- The probability of rolling a 1 when a die is rolled is $1/6$, meaning that one believe it.
- Probability is subjective.
- Parameter can be a random variable, and is updated using Bayes rule

$$\pi(\theta|\text{Data}) = \frac{\pi(\theta)P(\text{Data}|\theta)}{P(\text{Data})} \tag{10}$$

Example - binary classification

For some $f_\theta,\ \theta \in \Theta$, assume a model

$$y_i | x_i \sim Ber(f_\theta(x_i)). \tag{11}$$

Bayesian infer the probability distribution $\pi(\theta|D_n)$ (posterior distribution) of the parameter θ given the data D_n. From the probability distribution, one can suggest predictive distribution

$$p(y^{\text{new}}|x^{\text{new}}, D_n) = \int p(y^{\text{new}}|\theta, x^{\text{new}})\pi(\theta|D_n)d\theta. \tag{12}$$

That is, Bayesian can quantify uncertainty of the prediction.

Outline

Conclusion

- We investigated the statistical significance of supervised learning.
- In general, most machine learning models are closely related to statistical inference.
- Again, statistical inference is based on mathematical theories such as probability theory and functional analysis.
- So, please don't hate mathematics and statistics.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Thank You!