IT Learning

Annada Behera

NISER, Bhubaneswar

Advertisement

1. What is the theoretical minimum size of the data set?
2. What data points in the data set are important?
3. How to select the best model irrespective of the metric?

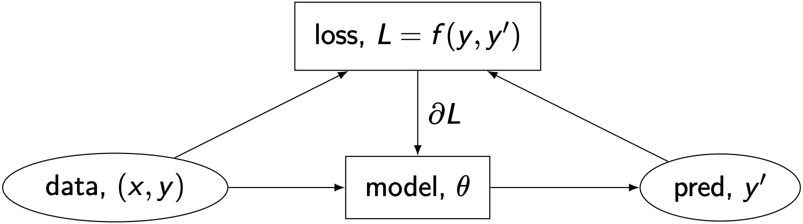
Advertisement

1. What is the theoretical minimum size of the data set?
2. What data points in the data set are important?
3. How to select the best model irrespective of the metric?

Bonus Model is generative for free! No need to train a GAN !

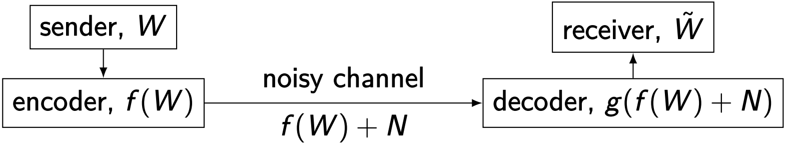
In short, learning is fast, less computationally expensive and theoretically with the best model.

Machine learning: overview



Why should the data be separated into x and y?

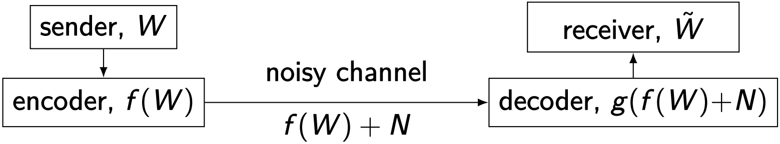
IT: Overview



Claude Shannon asked, "*How can we achieve perfect communication over an imperfect, noisy communication channel?*"

Unifying IT and ML

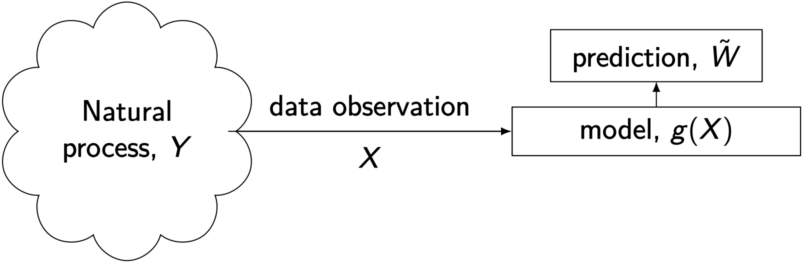
IT: How to predict the random variable W by observing the co-related random variable $Y = W + N$?



ML: How to the predict the true natural distribution of Y by sampling a co-related random variable X?

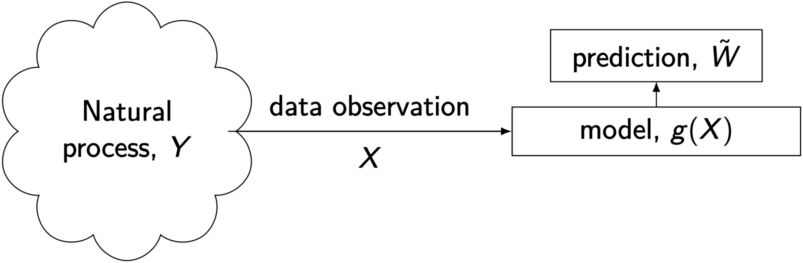
Unifying IT and ML

IT: How to predict the random variable W by observing the co-related random variable $Y = W + N$?



Unifying IT and ML

IT: How to predict the random variable W by observing the co-related random variable $Y = W + N$?



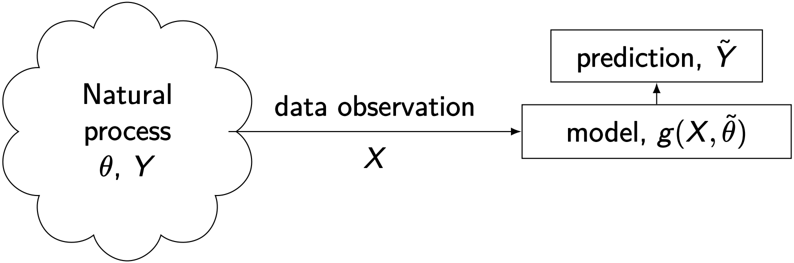
ML: How to the predict the true natural distribution of Y by sampling a co-related random variable X?

Probability theory: Definitions

Consider a process that we want to learn more about, i.e, predict and generate.

1. **Stochastic variables** These are the variables that take different values randomly.
2. **Non-stochastic variables**, (denoted by θ) are those which are fixed for a given system or process.
3. **Alphabet set**, $\mathcal{A}$, is a set of values that a given stochastic variable can take.
4. **Random variable**, X, is a mapping from the alphabet set, $\mathcal{A}$ to another measurable quantity.
5. **Distribution** of over a random variable X, $p(X)$ is defined as the probability of observing a given value of the variable.

Model fitting and prediction



In statistics (and ML), the process of model fitting, or the **maximum likelihood estimate** is done as,

$$\tilde{\theta} = \arg\max_{\theta} \Pr[\theta|X] \tag{1}$$

The techniques employed is mostly *gradient-based backprop*.

Probability theory: model fitting

The probability of the hypotheses θ, given the evidence X, is given by the Bayes rule,

$$p[\theta|X] = \frac{p[X|\theta] \cdot p[\theta]}{p[X]} \tag{2}$$

where

1. $p[\theta|X]$ is the **posterior** distribution after the model has seen the data.
2. $p[X|\theta]$ is the **likelihood** of seeing the data, if the hypotheses is true.
3. $p[\theta]$ is **prior** knowledge before seeing the data.
4. $p[X]$ is the probability of seeing the data.

Probability theory: model prediction and generation

In the context of probability theory, the model prediction can be computed by,

$$\Pr[x] = p(x \leq X \leq x + \epsilon|\theta) \tag{3}$$

As promised, the generative model is, as simple as computing the posterior predictive distribution,

$$p(\tilde{X}) = \int_{\Theta} d\theta \; p(\tilde{X}|\theta)p(\theta) \tag{4}$$

and sampling from it.

Probability theory: model prediction and generation

In the context of probability theory, the model prediction can be computed by,

$$\Pr[x] = p(x \leq X \leq x + \epsilon|\theta) \tag{3}$$

As promised, the generative model is, as simple as computing the posterior predictive distribution,

$$p(\tilde{X}) = \int_{\Theta} d\theta \; p(\tilde{X}|\theta)p(\theta) \tag{4}$$

and sampling from it.

But what is the prior distribution?

Bayes rule is subjective!

Bayes rule says,

$$p(\theta|X) = c \cdot p(X|\theta) \cdot p(\theta) \tag{5}$$

where, c is the **prior predictive** given as,

$$c = \frac{1}{p(X)} = \int_{\Theta} d\theta \; p(X|\theta) \cdot p(\theta) \tag{6}$$

And the likelihood is given by, our choice of our model.

- What model should we choose?
- What should be the prior?

Bayes rule is subjective!

Bayes rule says,

$$p(\theta|X) = c \cdot p(X|\theta) \cdot p(\theta) \tag{5}$$

where, c is the **prior predictive** given as,

$$c = \frac{1}{p(X)} = \int_{\Theta} d\theta \; p(X|\theta) \cdot p(\theta) \tag{6}$$

And the likelihood is given by, our choice of our model.

- What model should we choose?
- What should be the prior?

The answer was provided by **information theory**.

Shannon's information content

For any given distribution, $p(X)$ over a random variable, the **information content** or simply, *information* of any observation the random variable, $X = x$ is given by,

$$h(x) = \frac{1}{\log(x)} \tag{7}$$

where, the information is measured in *bits* or *shannon*.

Over the entire distribution, the expected information, called **entropy** is,

$$H(x) = \int_{x \in X(\mathcal{A})} dx\ p(x) \frac{1}{\log p(x)} \tag{8}$$

MaxEnt: Maximum Entropy Principle

- Without prior knowledge, use *principle of indifference* i.e, there must be no reason for suspecting one outcome over another.
- The distribution with maximum entropy has the least amount of information about the process, i.e, it is indifferent to any outcome.
- Use the distribution with MaxEnt for prior distribution.

Standby for demonstration !

Numerical approximation

If the posterior, $p(\theta|X)$ is in the same distribution family as the prior distribution $p(\theta)$, the prior and posterior are called **conjugate distributions** and the prior is called **conjugate prior**.

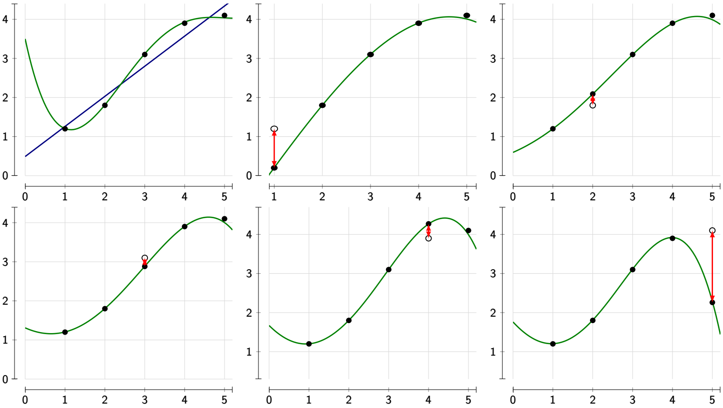
Example Beta distribution is a conjugate for binomial distribution.

$$B(r; N, f) = \binom{N}{r} f^r (1-f)^{N-r}, \quad B(m, n; f) = \frac{(1-f)^{m-1} f^n}{\mathrm{B}(m, n)} \tag{9}$$

- Grid approximation
- Laplace's (quadratic) approximation
- Metropolis-Hastings (MCMC with Gibbs sampling)
- Rejection sampling

Model selection

Leave-one-out cross-validation Drop every data point at a time and calculate the "out-of-sample" accuracy.



Widely Applicable Information Criteria

The KL divergence of two distribution, p and q is given as,

$$D(p||q) = \int_{-\infty}^{\infty} dx\ p(x) \log \frac{q(x)}{p(x)} \tag{10}$$

- Information criteria estimates the relative out-of-sample KL divergence.
- WAIC makes no assumptions about the posterior and it converges to the cross-validation in large samples.

Probabilitistic Programming Languages

- R
- OpenBUGS (Open-source WinBUGS)
- STAN (Andrew Gelman and team)
- PyMC (Uses, now deprecated, Theano backend)
- Pyro (Uses Facebook's Torch backend)
- Numpyro (Uses Google's JAX backend)

Questions?

Thank you!

References

- MacKay, David J.C. **Information Theory, Inference, and Learning Algorithms** Cambridge University Press, 2003
- McElreath, Richard **Statistical Rethinking, 2nd Ed.** CRC Press, 2020
- Sivia, D.S.; Skilling, J. **Data Analysis A Bayesian Tutorial, 2nd Ed.** Oxford University Press, 2006.
- Jaynes, E.T.; Bretthorst, G.L. **Probability Theory The Logic of Science**, Cambridge University Press, 2003.