Probability and Information Theory for ML

presented by Sung-Yub Kim

서울대학교 공과대학 산업공학과

January 6, 2016

[1] Kevin P. Murphy. Machine Learning - A Probabilistic Perspective *Adaptive Computation and Machine Learning*, MIT press, 2012.

[2] Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning *Computer Science and Intelligent Systems*, MIT Press, 2016.

Why Probability?

- Inherent stochasticity in the system being modeled.
  There exist really random factors and we need to model these factors in our model.
  e.g. Shuffle in card game, Customers or accidents come randomly
- Incomplete observability.
  We cannot observe system perfectly sometimes. Even the outcome is deterministic, if we don't know the outcome, we can think it stochastic.
  e.g. Monty Hall Game, Bayesian Statistics
- Incomplete modeling. For simplicity, we discard some information we already know. This case the information can be treated as stochastic.
  SimpleStochastic $\geq$ ComplexDeterministic
  e.g. Discretize space for robot.

One Issue: Frequentist VS Bayesian

- Frequentist
  "Probabilities represent long run frequencies of events."
  Computationally inexpensive relatively.
  Do not use subjective information.
  But there exist some limitations in performance.
- Bayesian
  "Probability is used to quatify our uncertainty about something or precisely degree of belief."
  Computationally expensive, since we need to compute all the distribution.
  Use subjective information.
  But the performance is better especially in high dimenstion.

Random Variable

- Random Experiment
  An experiment whose outcome cannot be predicted with certainty, before the experiment is run. In classical or frequency-based probability theory, we also assume that the experiment can be repeated indefinitely under essentially the same conditions.The repeatability assumption is important because the classical theory is concerned with the long-term behavior as the experiment is replicated. By contrast, subjective or belief-based probability theory is concerned with measures of belief about what will happen when we run the experiment. In this view, repeatability is a less crucial assumption.
- Sample Space
  The sample space of a random experiment is a set S that includes all possible outcomes of the experiment; the sample space plays the role of the universal set when modeling the experiment.
- Events
  Certain subsets of the sample space of an experiment are referred to as events.

Random Variable

- Random Variable
  Suppose again that we have a random experiment with sample space S. A function X from S into another set T is called a (T-valued) random variable. If sample space is a countable set, then we call this RV $discrete$. If sample space is an uncountable set, then we call this RV $continuous$.
- Probability measure
  A probability measure (or probability distribution) $\mathbb{P}$ for a random experiment is a real-valued function, defined on the collection of events, that satisifes the following axioms:
  a. $\mathbb{P}(A) \geq 0$ for every event A
  b. $\mathbb{P}(S) = 1$
  c. If $A_i : i \in I$is a countable, pairwise disjoint collection of events then

$$\mathbb{P}(\cup_{i \in I} A_i) = \sum_{i \in I} \mathbb{P}(A_i)$$

Probability Density Function

- Probability Density Function
  If X is a random variable, the probability density function of X is the function f on S that assigns probabilities to the points in S:
  a. $p(x) \geq 0, \forall x \in S$
  b. $\int_S p(x) = 1$
  c. $\int_A p(x) = \mathbb{P}(X \in A), \forall A \subseteq S$
  And pdf always exists by Radon-Nikodym Theorem. The value of pdf at each point means relative significancy of state.
- Probability Distribution
  A probability distribution of X is the function that assigns probabilities to the subset of S, namely $A \mapsto \mathbb{P}(X \in A)$ for $A \subseteq S$
  If we use parametrized distribution, then we can use the notation such as $X \sim U(a, b)$

Basic Rules

Now we use more simple notation $\mathbb{P}(X \in A)$ for $\mathbb{P}(A)$

- Sum Rule

$$\mathbb{P}(A + B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

- Product Rule

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

- Chain Rule
  By recursive product rule, we get

$$\mathbb{P}(X_{0:N-1}) = \mathbb{P}(X_0)\mathbb{P}(X_1|X_0)\cdots\mathbb{P}(X_{N-1}|X_{0:N-2})$$

- Marginal Probability

$$\mathbb{P}(A) = \int_A \int_Y p(x, y)dydx$$

- Conditional Probability

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\int_Y p(x, y)dy}$$

Bayes Rule

- Bayes Rule
  By Conditional probability, we get bayes rule

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x,y)}{\int_Y p(x,y)dy}$$

  We also use comprehendible notation like

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(H,D)}{\mathbb{P}(D)} = \frac{\mathbb{P}(H,D)}{\sum_H \mathbb{P}(H,D)} = \frac{\mathbb{P}(H)\mathbb{P}(D|H)}{\sum_H \mathbb{P}(H)\mathbb{P}(D|H)}$$

  And call $\mathbb{P}(H)$ prior, $\mathbb{P}(H|D)$ posterior, $\mathbb{P}(D|H)$ likelihood, and $\mathbb{P}(D)$ evidence. H means Hypothesis, D means Data.
- Generative Classifier
  Since we should know $\mathbb{P}(H,D)$ to make decision, we call generator use this concpet Generative Classifier.

Conditional Independence

- Unconditionally Independence
  We say RV X and Y are unconditionally independence if

$$p(x, y) = p(x)p(y), \forall x \in X, y \in Y$$

  This also means that $p(y) = p(y|x)$, $p(x) = p(x|y), \forall x \in X, y \in Y$. We symbolize unconditionally independence to $X \perp Y$
- Conditionally Independence
  We say RV X and Y are conditionally independent(CI) given Z if there exist function g and h suct that

$$p(x, y|z) = g(x, z)h(y, z), \forall x \in X, y \in Y, z \in Z$$

  We symbolize CI to $X \perp Y|Z$ In Probabilistic Graphical Model, we also use notation $X - Z - Y$, which means depedency between X and Y can be explained by Z

Expectation, Variance and Covariance

- Expectation
  Expectation means expected value of $f(x)$ when x is drawn form $f(x)$. We can get expectation of $f(x)$ by

$$\mathbb{E}_{X\sim P}[f(x)] = \int_X f(x)p(x)dx$$

- Expectations are linear (functional),

$$\mathbb{E}_X[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}_X[f(x)] + \beta\mathbb{E}_X[g(x)]$$

- Variance
  Variance means how much the values of a function of a random variable x vary as we sample different values of x from its distribution:

$$Var_X[f(x)] = \mathbb{E}_X[(f(x) - \mathbb{E}[f(x)])^2]$$

  The square root of the variation called standard deviation.
- Covariance
  Covariance gives us information about how two values are linearly related, as well as the scale of these variables:

$$Cov_{X,Y}(f(x), g(y)) = \mathbb{E}_{X,Y}[(f(x) - \mathbb{E}_X[f(x)])(g(y) - \mathbb{E}_Y[g(y)])^\top]]$$

Bernouli and Binomial Distributions

- Bernouli Distribution(Bin(θ))
  Bernouli Distribution has only one parameter θ which means the success probability of the trial. PMF of bernouli dist is shown like

$$Ber(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

- Binomial Distribution(Bin(n,θ))
  Binomial Distribution has two parameters n for number of trials, θ for success prob. PMF of binomial dist is shown like

$$Bin(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Multinomials and Multinouli Distributions

- Multinomial Distribution(Mu($\mathbf{x}|n,\theta$))
  Multinomial distribution is different from binomial with respect to dimension of ouput and θ. In binomial, k means the number of success. In multinomial each index of $\mathbf{x}$ means the number of state. Therefore we can see binomial as multinomial when the dimension of $\mathbf{x}$ and θ is 2.

$$Mu(\mathbf{x}|n,\theta) = \binom{n}{x_0,\ldots,x_{K-1}} \prod_{j=0}^{K-1} \theta_j^{x_j}$$

- Multinouli Distribution(Mu($\mathbf{x}|1,\theta$))
  Sometimes we are intersted in the special case of Multinomial when the n is 1 that is called Multinouli distribution:

$$Mu(\mathbf{x}|1,\theta) = \prod_{j=0}^{K-1} \theta_j^{\mathbb{I}(x_j=1)}$$

Beta and Dirichlet Distribution

- Beta Distribution(Beta($x|a,b$))

$$Beta(x|a,b) = \frac{1}{B(a,b)} x^a (1-x)^{b-1}$$

  Beta Distribution has two parameters a,b each counts how many occurs each classes. In Bayesian statistics, Beta is a conjugate prior of Bernouli and Binomial distribution.
- Dirichlet Distribution(Dir($\mathbf{x}|\alpha$))
  Before explain the Dirichlet, we need to define Probability Simplex like:

$$S_K = \{\mathbf{x} : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1\}$$

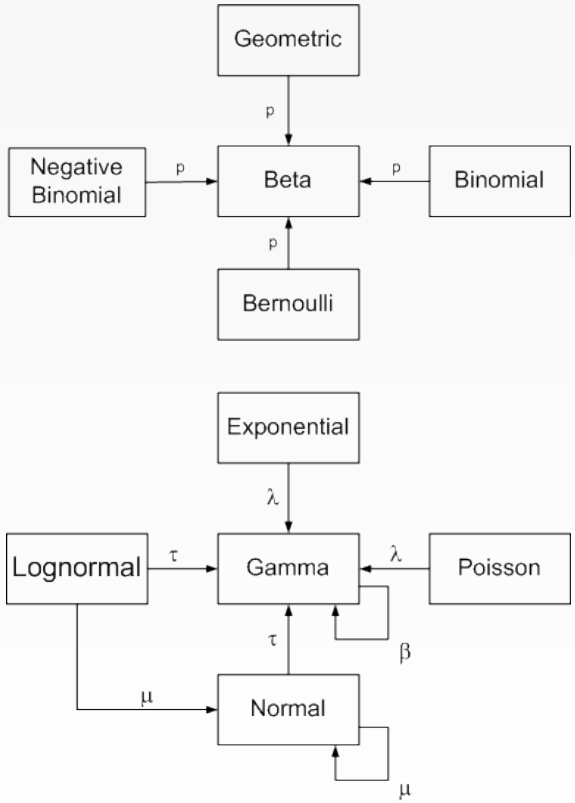
  Then Dirichlet Distribution is defined by

$$Dir(\mathbf{x}|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \mathbb{I}(\mathbf{x} \in S_K)$$

  Dirichlet Distribution has one parameter α which counts how many occurs each classes in its index. In Bayesian statistics, Dirichlet is a conjugate prior of and Multinouli and Multinomial distribution.

Conjugate Prior

- In Bayesian Statistics, we need to define prior for some parameteric distribution given parametric likelihood, since those distribution are comprehendible and computationally efficient. Since if we choose those priors, we can get posterior distribution in same distribtion, we call this prior "conjugate".



Poisson and Exponential Distribution

- Poisson Distribution(Poi($x|\lambda$))
  Poisson distribution has only one parameter λ which means the average occurence number. This distribution let us know the probability of how many time it will occur.

$$Poi(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$$

- Exponential Distribution(Exp($x|\lambda$))
  Exponentail distribution also has only one parameter λ that is same as poisson. But this distribution let us know the probability of how much time do we need to wait for occurence.

$$Exp(x|\lambda) = \mathbb{I}_{x\geq 0}\lambda e^{-\lambda x}$$

Gaussian Distribution

- Gaussian Distribution($\mathcal{N}(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\beta^{-1})$)

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

$$\mathcal{N}(x|\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} exp(-\frac{\beta}{2}(x-\mu)^2)$$

  μ means average, σ^2 means variance, β means precision which is the reciprocal of variance.
  Gaussian Distribution can be a sensible choice for many applications.

- The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behavior.
- Second, out of all possible probability distributions with same mean and variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers.(Maximum Entropy Principle) We can thus think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model.

Multivariate Gaussian Distribution

In many case, we need to consider many gaussian variables at once, therefore we need tool to use it.Multivariate Gaussian Distribution($\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \mathcal{N}(\mathbf{x}|\mu,\beta^{-1})$)

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \sqrt{\frac{1}{(2\pi)^n det(\Sigma)}} exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu))$$

$$\mathcal{N}(\mathbf{x}|\mu,\beta^{-1}) = \sqrt{\frac{det(\beta)}{(2\pi)^n}} exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \beta(\mathbf{x}-\mu))$$

It is almost same for univariate version, but σ^2 or β which are positive scalr become PSD matrix Σ or .(But the reciprocal relationship is maintain, $\beta = \Sigma^{-1}$)

Laplace Distribution

- Laplace Distribution(Lap($x|\mu, \gamma$))
  We can observe that the formula of gaussian distribution measure how far the point from μ. We can formulate similar definition, but this distribution has a sharp peak at μ

$$Lap(x|\mu, \gamma) = \frac{1}{2\gamma} exp(-\frac{|x-\mu|}{\gamma})$$

  This distribution has robustness to outlier. It also put more probability density at 0 than gaussian. Therefore it has some sparsity property that is useful many way.

Dirac and Empirical Distriubtion

- Dirac Distribution($\delta_\mu(x)$)
  Dirac Distribution is not a function but generalized function which is defined in terms of limit of Gaussian Distribution.

$$\delta_\mu(x) = \lim_{\beta \to \infty} \mathcal{N}(x|\mu, \beta^{-1})$$

  And it means these

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

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

- Empirical Distribution
  Empirical Distribution is defined by mixture distribution of Dirac distribution:

$$\hat{p}(x) = \frac{1}{m} \sum_{i=0}^{m-1} \delta_{x_i}(x)$$

  We can view the empirical distribution formed from a dataset of training examples as specifying the distribution that we sample from when we train a model on this dataset. Another important perspective on the empirical distribution is that it is the probability density that maximizes the likelihood of the training data

Mixture of Distribution

- Make complex distribution by combining simpler distribution!
  A mixture distribution is made up of several component distributions. On each trial, the choice of which component distribution generates the sample is determined by sampling a component identity from a multinoulli distribution:

$$\mathbb{P}(x) = \sum_i \mathbb{P}(c = i)\mathbb{P}(x|c = i)$$

The mixture model allows us to briey glimpse a concept that will be of paramount importance later—the latent variable(In this case c is the latent variable).
A very powerful and common type of mixture model is the Gaussian mixture model, in which the components $p(x|c = i)$ are Gaussians. In this case, $\alpha_i = \mathcal{P}(c = i)$ are prior, and $\mathcal{P}(c = i|x)$ are posterior.
A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specic, non-zero amount of error by a Gaussian mixture model with enough components.

Sigmoid and Softplus Function

- Sigmoid Function

$$\sigma(x) = \frac{1}{1 + \exp(x)}$$

Sigmoid Function is commonly used to produce the ϕ parameter of a Bernouli distribution because its range is (0,1), which lies within the valid range of values for the ϕ paramter.

- Softplus Function

$$\zeta(x) = \log(1 + \exp(x))$$

Softplus can be useful for producing the β or σ parameter of a normal distribution because its range is $(0, \infty)$. The name of softplus fnuction comes from the fact that it is a softened version of

$$x^+ = \max(0, x)$$

Sigmoid and Softplus Function

- Useful Properties

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)}$$

  Therefore, it can be generalized in Multinouli distribution by $\sigma(\mathbf{x}) = (\frac{\exp(x_i)}{\sum_i \exp(x_i)})_i$

- Derivate property of sigmoid and softplus

$$\frac{d\zeta(x)}{dx} = \sigma(x), \zeta(x) = \int_{-\infty}^{x} \sigma(y)dy, \frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$$

- Similarity with plus function and softplus

$$\zeta(x) - \zeta(-x) = x, 1 - \sigma(x) = \sigma(-x)$$

- Another way for softplus to sigmoid

$$\log(\sigma(x)) = -\zeta(-x)$$

- Inverse property

$$\forall x \in (0, 1), \sigma^{-1}(x) = \log(\frac{x}{1 - x}), \forall x > 0, \zeta^{-1}(x) = \log(\exp(x) - 1)$$

Transformations of Random Variables

- Linear Transformation
  Suppose $x \mapsto f(x)$ is a linear trnasformation:

$$y = f(x) = Ax + b$$

  Since Expectation is linear (functional), we can get:

$$\mathbb{E}[y] = \mathbb{E}[Ax + b] = A\mathbb{E}[x] + b$$

$$Cov[y] = Cov[Ax + b] = \mathbb{E}[(Ax - A\mathbb{E}[x])(Ax - A\mathbb{E}[x])^\top] = A\Sigma A^\top$$

- General Transformation and Change of Variable
  For genral transformation f

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq f(x))$$

$$p(y) = \frac{d}{dy} P_Y(y) = \frac{d}{dy} P_X(f^{-1}(y)) = \frac{dx}{dy}\frac{d}{dx} P_X(x) = \frac{dx}{dy} p(x)$$

  Since this transformation is limited in volume, we note like

$$p(y) == |\frac{dx}{dy}| p(x)$$

  In multivariate case, we generalize this and get:

$$p(\mathbf{y}) = |\det J_{\mathbf{y}\to\mathbf{x}}| p(\mathbf{x})$$

Entropy

- Principle of quantifing Infotmation
  1. Likely events should have low information content. Less likely events should have higher information content
  2. Independent events should have additive information.
- Self-Information
  By above principle, we can define self-information like:

$$I(x) = -\log P(x)$$

  If base of $\log$ is ϵ, then we call this information nats. If base of $\log$ is 2, then we call this information bits.
- Shannon Entropy
  We can quantify uncertainty of whole distribution like:

$$H(x) = \mathbb{E}_{x\sim P}[I(x)] = -\mathbb{E}_{x\sim P}[\log P(x)]$$

  We also denote this entropy by $H(P)$.
  Shannon entropy means expected amount of information in an event drawn from that distribution.

Cross Entropy

- Kullbeck-Leibler(KL) Divergence
  We sometimes want to know how different two distributions are. In this case, we use KL divergence defined by:

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

  And it means extra amount of information needed to send a message containing symbols drawn from probability distribution P, when we use a code that was designed to minimize the length of messages drawn from probability distribution .

- Cross-Entropy
  Cross-entropy which means the average number of information needed to encode data coming from a source with distripution P when we use model q to define our codebook.(not extra)
  We can define cross-entropy like:

$$H(P,Q) = H(P) + KL(P\|Q) = -\mathbb{E}_{x\sim P}[\log Q(x)]$$

GM

- Why we use?
  Using a single function to describe the entire joint probability distribution can be very inecient.
  Instead of using a single function to represent a probability distribution, we can split a probability distribution into many factors that we multiply together like this:

$$p(a, b, c) = p(a)p(b|a)p(c|a, b)$$

  In this way, we can can greatly reduce the cost of representing a distribution if we are able to find a factorization into distributions over fewer variables.
- GM
  We can describe these kinds of factorizations using graphs. Here we use the word "graph" in the sense of graph theory: a set of vertices that may be connected to each other with edges. When we represent the factorization of a probability distribution with a graph, we call it a structured probabilistic model or graphical model.

Directed and Undirected Graph

- Directed model
  Directed models use graphs with directed edges, and they represent factoriza- tions into conditional probability distributions, as in the example above. In this model, we can factorize the distribution like:

$$p(\mathbf{x}) = \prod_i p(x_i | P_a\mathcal{G}(x_i))$$

- Undirected model
  Undirected models use graphs with undirected edges, and they represent fac- torizations into a set of functions; unlike in the directed case, these functions are usually not probability distributions of any kind. Any set of nodes that are all connected to each other in $\mathcal{G}$ is called a clique. Each clique $\mathcal{C}^{(i)}$ is an undirected model is associated with a factor $\phi^{(i)}(\mathcal{C}^{(i)})$ These factors are just functions, not probability distributions. The output of each factor must be non-negative, but there is no constraint that the factor must sum or integrate to 1 like a probability distribution.
  We therefore divide by a normalizing constant Z, dened to be the sum or integral over all states of the product of the ϕ functions, in order to obtain a normalized probability distribution:

$$p(x) = \frac{1}{Z} \prod_i \phi^{(i)}(\mathcal{C}^{(i)})$$