Probability Distributions for ML

Sung-Yub Kim

Dept of IE, Seoul National University

January 29, 2017

- Bishop, C. M. Pattern Recognition and Machine Learning *Information Science and Statistics*, Springer, 2006.
- Kevin P. Murphy. Machine Learning - A Probabilistic Perspective *Adaptive Computation and Machine Learning*, MIT press, 2012.
- Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning *Computer Science and Intelligent Systems*, MIT Press, 2016.

- Purpose: Density Estimation
- Assumption: Data Points are independent and identically distributed.(i.i.d)
- Parametric and Nonparametric
  Parametric estimations are more intuitive but has very strong assumption.
  Nonparametric estimation also has some parameters, but they control model complexity.

- Bernouli Distribution(Ber(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

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

- Likelihood of Data
  By i.i.d assumption, we get

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \tag{1}$$

- Log-likelihood of Data
  Take logarithm, we get

$$\ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\} \tag{2}$$

- MLE
  Since maximizer is stationary point, we get

$$\mu_{ML} := \hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{3}$$

- Prior Distribution
  The weak point of MLE is you can be overfitted to data. To overcome this deficiency, we need to make some prior distribution.
  But same time our prior distribution need to has **a simple interpretation** and **useful analytical properties**.
- Conjugate Prior
  Conjugate prior for a likelihood is a prior distribution which your prior and posterior distribution are same given your likelihood.
  In this case, we need to make our prior proportional to powers of μ and $(1-\mu)$. Therefore, we choose Beta Distribution

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{4}$$

Beta Distribution has two parameters a,b each counts how many occurs each classes(**effective number of observations**). Also we can easily valid that posterior is also beta distribution.

- Posterior Distribution
  By some calculation,

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \quad (5)$$

  where m,l are observed data.

- Bayesian Inference
  Now we can make some bayesian inference on binary variables. We want to know

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu = \int_0^1 \mu p(\mu|\mathcal{D})d\mu = \mathbb{E}[\mu|\mathcal{D}] \quad (6)$$

  Therefore we get

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b} \quad (7)$$

  If observed data(m,l) are sufficiently big, its asymptotic property is identical to MLE, and this property is very general.

Since

$$\mathbb{E}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]] \tag{8}$$

we know that poseterior mean of θ, averaged over the distribution generating the data, is equal to the prior mean of θ.
Also since

$$Var_\theta[\theta] = \mathbb{E}_\mathcal{D}[Var_\theta[\theta|\mathcal{D}]] + Var_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]] \tag{9}$$

We know that on average, the posterior variance of θ is smaller than the prior variance.

- 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:

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

- Likelihood of Data
  By i.i.d assumption, we get

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{\sum_n x_{nk}} = \prod_{k=1}^{K}\mu_k^{m_k} \tag{10}$$

  where $m_k = \sum_n x_{nk}$ (**sufficient statistics**)
- Log-likelihood of Data
  Take logarithm, we get

$$\ln p(D|\mu) = \sum_{k=1}^{K} m_k \ln \mu_k \tag{11}$$

- MLE
  Therefore, we need to solve following optimization problem for MLE

$$\max\{\sum_{k=1}^{K} m_k \ln \mu_k | \sum_{k=1}^{K} \mu_k = 1\} \tag{12}$$

- MLE(cont.)

  We already know that Lagrangian stationaty point is a necessary condition for constrained optimization problem. Therefore,

$$\nabla_{\mu}\mathcal{L}(\mu;\lambda) = 0, \nabla_{\lambda}\mathcal{L}(\mu;\lambda) = 0 \tag{13}$$

  where

$$\mathcal{L}(\mu;\lambda) = \sum_{k=1}^{K} m_k \ln \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1) \tag{14}$$

  Therefore, we get

$$\mu_k^{ML} = \frac{m_k}{N} \tag{15}$$

- Dirichlet Distribution
  By the same intuition in Beta distribution, we can get conjugate prior for Multinouli
  $$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{16}$$
  where $\alpha_0 = \sum_k \alpha_k$
- Bayesian Inference
  By the same argument in binomial, we can get posterior probability
  $$p(\mu|\mathcal{D},\alpha) = Dir(\mu|\alpha + m) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \tag{17}$$

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

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \tag{18}$$

$$\mathcal{N}(x|\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} exp(-\frac{\beta}{2}(x-\mu)^2) \tag{19}$$

- Multivariate Gaussian Distribution($\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \mathcal{N}(\mathbf{x}|\mu,\beta^{-1})$)

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} det(\Sigma)^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)) \tag{20}$$

$$\mathcal{N}(\mathbf{x}|\mu,\beta^{-1}) = \frac{1}{(2\pi)^{\frac{D}{2}} det(\Sigma)^{\frac{1}{2}}} exp(-\frac{1}{2}(\mathbf{x}-\mu)^\top \beta(\mathbf{x}-\mu)) \tag{21}$$

- Mahalanobis Distance
  By EVD, we can get

$$\Delta^2 = (x-\mu)^\top \Sigma^{-1}(x-\mu) = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \tag{22}$$

where $y_i = u_i^\top (x-\mu)$

- Change of Variable in Gaussian
  By above, we can get

$$p(y) = p(x)|J_{y\rightarrow x}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{\frac{1}{2}}} \exp\{-\frac{y_j^2}{2\lambda_j}\} \tag{23}$$

which means product of D independent univariate Gaussian Distribution.

- First and Second Moment of Gaussian
  By using above, we can get

$$\mathbb{E}[x] = \mu, \mathbb{E}[xx^\top] = \mu\mu^\top + \Sigma \tag{24}$$

- Limitations of Gaussian and Solutions
  There are two main limitations for Gaussian.
  First, we have to infer so many covariance parameters.
  Second, we cannot represent multi-modal ditriubtions. Therefore, we define some auxilarily concepts.
- Diagonal Covariance

$$\Sigma = diag(s^2) \tag{25}$$

- Isotropic Covariance

$$\Sigma = \sigma^2 I \tag{26}$$

- Mixture Model

$$p(x) = \sum_{k=1}^{K} \pi_k p(x|\pi_k) \tag{27}$$

- Partitions of Mahalanobis distance
  First, partition the covariance matrix and precision matrix.

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}, \Sigma^{-1} = \Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \tag{28}$$

  where aa, bb are symmetric and ab and ba are conjugate transpose.
  Now, partition the Mahalanobis distance.

$$(x-\mu)^\top \Sigma^{-1}(x-\mu)$$

$$\begin{aligned} &= (x_a-\mu)^\top \Sigma_{aa}^{-1}(x_a-\mu) + (x_a-\mu)^\top \Sigma_{ab}^{-1}(x_b-\mu) \\ &\quad + (x_b-\mu)^\top \Sigma_{ba}^{-1}(x_a-\mu) + (x_b-\mu)^\top \Sigma_{bb}^{-1}(x_b-\mu) (29) \end{aligned}$$

- Schur Complement
  Like gaussian elimination, we can use some block matrix elimination by **Schur Complement**

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{bmatrix} \tag{30}$$

  where $M = (A - BD^{-1}C)^{-1}$

- Schur Complement(cont.)
  Therefore, we get

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \tag{31}$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1} \tag{32}$$

- Conditional Distribution
  Therefore, we get

$$x_a|x_b \sim \mathcal{N}(x|\mu_{a|b}, \Sigma_{a|b}) \tag{33}$$

where

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - x_a) \tag{34}$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \tag{35}$$

- Marginal Distribution
  Removing x_b by integrating, we can get marginal distribution of x_a

$$p(x_a) = -\frac{1}{2}x_a^\top(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}\Lambda_{ba})x_a + x_a^\top(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}\Lambda_{ba})\mu_a + const \tag{36}$$

Therefore, we get

$$x_a \sim \mathcal{N}(x|\mu_a, \Sigma_{aa}) \tag{37}$$

Given a marginal Gaussian for x and a conditional Gaussian for y given x in the form

$$x \sim \mathcal{N}(x|\mu, \Lambda^{-1}) \tag{38}$$

$$y|x \sim \mathcal{N}(y|Ax + b, L^{-1}) \tag{39}$$

Then we can get marginal distribution of y and the conditional distribution of x given y are given by

$$y \sim \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^\top) \tag{40}$$

$$x|y \sim \mathcal{N}(x|\Sigma\{A^\top L(y - b) + A\mu\}, \Sigma) \tag{41}$$

where

$$\Sigma = (\Lambda + A^\top LA)^{-1} \tag{42}$$

- Log-likelihood for data
  By same argument in categorical data, we can get log-likelihood for Gaussian

$$\ln p(D|\mu, \Sigma) = -\frac{ND}{2}\ln 2\pi - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^\top \Sigma^{-1}(x_n - \mu) \quad (43)$$

  and this log-likelihood depends only on these quantities called **Sufficient Statistics**

$$\sum_{n=1}^{N} x_n, \sum_{n=1}^{N} x_n x_n^\top \quad (44)$$

- MLE for Gaussian
  Since MLE is a maximizer for log-likelihood, we can get

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad (45)$$

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(x_n - \mu_{ML})^\top \quad (46)$$

- Sequential estimation
  Since we get MLE for gaussian analytically, we can do this sequentially like

$$\mu_{ML}^{N} = \mu_{ML}^{N-1} + \frac{1}{N}(x_N - \mu_{ML}^{N-1}) \tag{47}$$

- Robbins-Monro Algorithm
  By same intuition, we can generalize sequential learning. Robbins-Monro algorithm gives us root θ such that $f(\theta) = \mathbb{E}[z|\theta] = 0$. The iterate process of RM algorithm can be represented by

$$\theta^{N} = \theta^{N-1} - a_{N-1} z(\theta^{N-1}) \tag{48}$$

where $z(\theta^{N-1})$ means observed value of z when θ takes the value θ^{N-1} and a_N is an sequence satisfy

$$\lim_{N\to\infty} a_N = 0, \sum_{N=1}^{\infty} a_N = \infty, \sum_{N=1}^{\infty} a_N < \infty \tag{49}$$

- Generalized Sequential Learning
  We can apply RM algorithm for sequential learning. In this case, our $f(\theta)$ is a gradient of log-likelihood function. Therefore, we can get

$$z(\theta) = -\frac{\partial}{\partial\theta} \ln p(x|\theta) \tag{50}$$

  In Gaussian case, we put a_N to σ^2/N.

- Bayesian Inference for mean given variance
  Since gaussian likelihood takes the form of the exponential of a quadratic form in μ, we can choose a prior also Gaussian. Therefore, if we choose

$$\mu \sim \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{51}$$

  for prior, we get following for posterior

$$\mu|\mathcal{D} \sim \mathcal{N}(\mu|\mu_N, \sigma_N^2) \tag{52}$$

  where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}, \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \tag{53}$$

- Bayesian Inference for mean given variance(cont.)
  1. Posterior mean compromises between the priot and the MLE.
  2. Precision is given by the precision of the prior plus one contribution of the data precision from each of the observed data.
  3. If we take $\sigma_0^2 \to \infty$ then the posterior mean reduces to the MLE.
- Bayesian Inference for variance given mean
  Since gaussian likelihood takes the form of proportional to the product of a power of precision and the exponential of a linear function of precision. We choose gamma distribution which is defined by

$$Gam(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a}{}_0 \lambda^{a_0 - 1} \exp(-b_0 \lambda) \tag{54}$$

  Then we can get posterior

$$\lambda|\mathcal{D} \sim Gam(\lambda|a_N, b_N) \tag{55}$$

  where

$$a_N = a_0 + \frac{N}{2}, b_N = b_0 + \frac{N}{2}\sigma_{ML}^2 \tag{56}$$

- Bayesian Inference for variance given mean(cont.)
  1. We can interpret the parameter $2a_0$ effective prior observations for number of data. 2. We can interpret the parameter b_0/a_0 effective prior observations for variance.
- Bayesian Inference for no data
  By apply same argument on mean and variance, we can get prior

$$p(\mu, \lambda) \sim \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1}) Gam(\lambda|a, b) \tag{57}$$

where

$$\mu_0 = c/\beta, a = 1 + \beta/2, b = d - c^2/2\beta \tag{58}$$

Note that precision of μ is a linear function of λ
For Multivariate case, we can similarly get prior

$$p(\mu, \Lambda|\mu_0, \beta, W, \nu) = \mathcal{N}(\mu|\mu_0, (\beta\Lambda)^{-1})\mathcal{W}(\Lambda|W, \nu) \tag{59}$$

where $\mathcal{W}$ is Wishart distribution.

- Univariate t-distribution
  If we integrate out the precision given that our prior for precision is Gamma, we get t-distribution.

$$St(x|\mu,\lambda,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}(\frac{\lambda}{\pi\nu})^{1/2}[1+\frac{\lambda(x-\mu)^2}{\nu}]^{-\nu/2-1/2} \tag{60}$$

  where $\nu = 2a$(**degrees of freedom**) and $\lambda = a/b$.
  We can think t-dstribution as an **infinite mixture of Gaussians**.
  Since t-distribution has fat tail(than Gaussian), we can obtain more robust model when we estimate.

- Multivariate t-distribution
  We also can get multivariate case of infinite mixture of Gaussians, then we get multivariate t-distribution

$$St(x|\mu,\Lambda,\nu) = \frac{\Gamma(\nu/2+D/2)}{\Gamma(\nu/2)}(\frac{\Lambda^{1/2}}{(\pi\nu)^{D/2}})[1+\frac{\Delta^2}{\nu}]^{-\nu/2-D/2} \tag{61}$$

- The Exponential Family
  The exponential family of distributions over x, given parameters η, is defined to be the set of distributions of the form

$$p(x|\eta) = g(\eta)h(x)\exp\{\eta^\top u(x)\} \tag{62}$$

  where η is **natural parameters** of the distribution, and $u(x)$ is a function of x.
  The fnuction $g(\eta)$ can be interpereted as the normalization factor.

- Logistic Sigmoid
  In case of bernouli distribution, our parameter is μ, although our natural parameter is η. Those two parameter can be connected by following
  $$\eta = \ln(\frac{\mu}{1-\mu}), \mu := \sigma(\eta) = \frac{\exp(\mu)}{1+\exp(\mu)} \tag{63}$$
  And we call this $\sigma(\eta)$ **sigmoid function**.
- Softmax function
  By same argument, we can find some realtionship between our parameter and natural parameter. That is **Softmax function**.
  $$\mu_k = \frac{\exp(\eta_k)}{\sum_{j=1}^{K} \exp(\eta_j)} \tag{64}$$
  Note that in this case, $u(x) = 1, h(x) = 1, g(x) = (\sum_{j=1}^{K} \exp(\eta_j))^{-1}$

- Gaussian
  Gaussian also can be interpreted as the exponential family by

$$u(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \tag{65}$$

$$\eta = \begin{bmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \tag{66}$$

$$g(\eta) = (-2\eta_2)^{1/2} \exp(\frac{\eta_1^2}{4\eta_2}) \tag{67}$$

- Problem of estimating the natural parameter
  We can generalize the argument in MLE in other cases.
  First, we consider the log-likelihood of the data.

$$\ln p(\mathcal{D}|\eta) = \sum_{n=1}^{N} h(x_n) + N \ln g(\eta) + \eta^\top \sum_{n=1}^{N} u(x_n) \tag{68}$$

Next, we need to find the stationary point of the log-likelihood.

$$N\nabla_\eta \ln g(\eta) + \sum_{n=1}^{N} u(x_n) = 0 \tag{69}$$

Therfore, we get MLE

$$-\nabla_\eta \ln g(\eta) = \frac{1}{N} \sum_{n=1}^{N} u(x_n) \tag{70}$$

We see that the solution for the MLE depedns on the data only through $\sigma_n u(x_n)$, which is therefore called the **sufficient statistic** of the exponential family.

- Conjugate prior
  For any member of the exponential family, there exists a conjugate prior that can be written in the form

$$p(\eta|\chi,\nu) = f(\chi,\nu)g(\eta)^{\nu}\exp\{\nu\eta^{\top}\chi\} \tag{71}$$

  where $f(\chi,\nu)$ is a normalization factor, and $g(\eta)$ is the same function as the exponential family.

- Posterior distribution
  If we choose prior as conjugate prior, we get

$$p(\eta|\mathcal{D},\chi,\nu) \propto g(\eta)^{\nu+N}\exp\{\eta^{\top}(\sum_{n=1}^{N}u(x_n)+\nu\chi)\} \tag{72}$$

  Therefore, we see that the parameter ν can be interpreted as the **effective number of pseudo-observations** in the prior, each of which has a value for the sufficient statistics $u(x)$ given by χ.

- Noninformative Priors
  We may seek a form of prior distribution, called a **noninformative prior**, which is intended to have as little influence on the posterior distribution as possible.
- Generalizations of Noninformative priors
  It leads to two generalizations, namely the principle of transformation groups as in the Jeffreys prior, and the principle of maximum entropy.

- Histogram Technique
  Standard histograms simply partition x into distinct bins of width Δ_i and then count the number ni of observations of x falling in bin i. In order to turn this count into a normalized probability density, we simply divide by the total number N of observations and by the width Δ_i of the bins to obtain probability values for each bin given by

$$p_i = \frac{n_i}{N\Delta_i} \tag{73}$$

- Limitations of Hitogram
  The estimated density has discontinuities that are due to the bin edges rather than any property of the underlying distribution that generated the data.
  Histogram approach also sacling with dimensionality.
- Lessons of Histogram
  First, to estimate the probability density at a particular location, we should consider the data points that lie within some local neighbourhood of that point.
  Second, the value of the smoothing parameter should be neither too large nor too small in order to obtain good results.

- Motivation

For large N, the bernouli trial that data point fall within small region *mathcalR* will be sharply peaked around the mean and so

$$K \simeq NP \tag{74}$$

If, however, we also assume that the region $\mathcal{R}$ is sufficiently small that the probability density p(x) is roughlt over the region, then we have

$$P \simeq p(x)V \tag{75}$$

where V is the volume of $\mathcal{R}$. Therefore,

$$p(x) = \frac{K}{NV} \tag{76}$$

Note that in our assumption, $\mathcal{R}$ is sufficiently small tha the density is approximately constant over the region and the yet sufficiently large that the number K of points falling inside the region is sufficient for the binomial distribution to be sharply peaked.

- Kernel Density Estimation(KDE)
  If we fix V and determine K from the data, we use kernel approach. For instance, we fix V to 1 and count the data point by following function

$$k(u) = \begin{cases} 1, if|u_i| \leq 1/2, i = 1, \cdots, D, \\ 0, otherwise \end{cases} \tag{77}$$

  which called **Parzen window** In this case, we can use this by

$$K = \sum_{n=1}^{N} k(\frac{x - x_n}{h}) \tag{78}$$

  and it leads density function

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k(\frac{x - x_n}{h}) \tag{79}$$

  We can also use another kernel like Gaussian kernel. If we do so, then we get

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{D/2}} \exp\{-\frac{\|x - x_n\|}{2h^2}\} \tag{80}$$

- Limitation of KDE
  One of the difficulties with the kernel approach to density estimation is that the parameter h governing the kernel width is fixed for all kernels. In regions of high data density, a large value of h may lead to over-smoothing and in lower data density, a small value of h may lead to overfitting. Thus the optimal choice for h may be dependent on location within data space.
- Nereat-Neighbor(NN)
  Therefore we consider a fixing K and use the data to find an appropriate V and we call this method K-NN methods.
  In this case, the value of K governs the degree of smoothing and we need to optimizae(hyper-parameter optimize) K.
- Erro of KNN
  Note that for sufficiently big N, the error rate is never more than twice the minimum achievable error rate of an optimal classifier.