Pattern Recognition and Machine Learning
Chapter 2: Probability Distributions

Cécile Amblard   Alex Kläser   Jakob Verbeek

October 11, 2007

Probability Distributions: General

- **Density Estimation:** given a finite set $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of observations, find distribution $p(\mathbf{x})$ of $\mathbf{x}$
  - **Frequentist's Way:** chose specific parameter values by optimizing criterion (e.g., likelihood)
  - **Bayesian Way:** prior distribution over parameters, compute posterior distribution with Bayes' rule
- **Conjugate Prior:** leads to a posterior distribution of the same functional form as the prior (makes life a lot easier :)

Binary Variables: Frequentist's Way

Given a binary random variable $x \in \{0, 1\}$ (tossing a coin) with

$$p(x = 1|\mu) = \mu, \quad p(x = 0|\mu) = 1 - \mu. \tag{2.1}$$

$p(x)$ can be described by the *Bernoulli distribution*:

$$\mathrm{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}. \tag{2.2}$$

The *maximum likelihood* estimate for μ is:

$$\mu^{\mathrm{ML}} = \frac{m}{N} \quad \text{with} \quad m = (\#\text{observations of } x = 1) \tag{2.8}$$

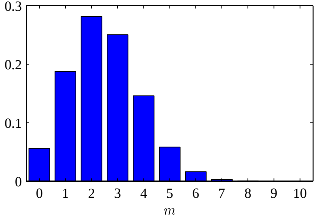
Yet this can lead to overfitting (especially for small N), e.g., $N = m = 3$ yields $\mu^{\mathrm{ML}} = 1$!

Binary Variables: Bayesian Way (1)

The *binomial distribution* describes the number m of observations of $x = 1$ out of a data set of size N:

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \quad (2.9)$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \quad (2.10)$$



Binary Variables: Bayesian Way (2)

For a Bayesian treatment, we take the *beta distribution* as conjugate prior:

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1)} \tag{2.13}$$

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}\mathrm{d}u$$

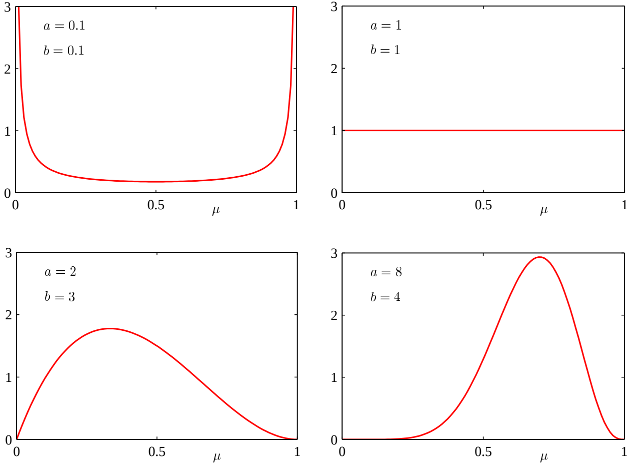
(The gamma function extends the factorial to real numbers, i.e., $\Gamma(n) = (n-1)!$.) Mean and variance are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.15}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.16}$$

Binary Variables: Beta Distribution

Some plots of the beta distribution:



Binary Variables: Bayesian Way (3)

Multiplying the binomial likelihood function (2.9) and the beta prior (2.13), the posterior is a beta distribution and has the form:

$$\begin{aligned} p(\mu|m,l,a,b) &\propto \mathrm{Bin}(m,l|\mu)\mathrm{Beta}(\mu|a,b) \\ &\propto \mu^{m+a-1}(1-\mu)^{l+b-1} \end{aligned} \tag{2.17}$$

with $l = N - m$.

- Simple interpretation of hyperparameters a and b as effective number of observations of $x = 1$ and $x = 0$ (a priori)
- As we observe new data, a and b are updated
- As $N \to \infty$, the variance (uncertainty) decreases and the mean converges to the ML estimate

Multinomial Variables: Frequentist's Way

A random variable with K mutually exclusive states can be represented as a K dimensional vector $\mathbf{x}$ with $x_k = 1$ and $x_{i \neq k} = 0$. The *Bernoulli distribution* can be generalized to

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{2.26}$$

with $\sum_k \mu_k = 1$. For a data set $\mathcal{D}$ with N independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the corresponding likelihood function takes the form

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k} \tag{2.29}$$

The *maximum likelihood* estimate for $\boldsymbol{\mu}$ is:

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N} \tag{2.33}$$

Multinomial Variables: Bayesian Way (1)

The *multinomial distribution* is a joint distribution of the parameters $m_1, \ldots, m_K$, conditioned on $\boldsymbol{\mu}$ and N:

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \quad (2.34)$$

$$\binom{N}{m_1 m_2 \ldots m_K} \equiv \frac{N!}{m_1! m_2! \ldots m_K!} \quad (2.35)$$

where the variables m_k are subject to the constraint:

$$\sum_{k=1}^{K} m_k = N \quad (2.36)$$

Multinomial Variables: Bayesian Way (2)

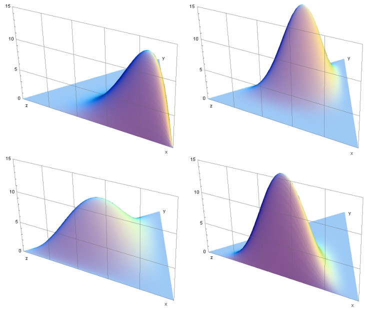
For a Bayesian treatment, the *Dirichlet distribution* can be taken as conjugate prior:

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.38}$$

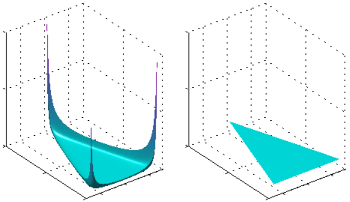
with $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.

Multinomial Variables: Dirichlet Distribution

Some plots of a Dirichlet distribution over 3 variables:



Dirichlet distribution with values (clockwise from top left): $\alpha = (6,2,2),(3,7,5),(6,2,6),(2,3,4)$.



Dirichlet distribution with values (from left to right): $\alpha = (0.1,0.1,0.1),(1,1,1)$.

Multinomial Variables: Bayesian Way (3)

Multiplying the prior (2.38) by the likelihood function (2.34) yields the posterior:

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \quad (2.40)$$

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \quad (2.41)$$

with $\mathbf{m} = (m_1, \ldots, m_K)^\top$. Similarly to the binomial distribution with its beta prior, α_k can be interpreted as effective number of observations of $x_k = 1$ (a priori).

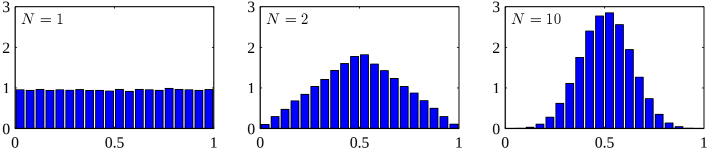
The gaussian distribution

The gaussian law of a D dimensional vector $\mathbf{x}$ is:

$$N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\} \quad (2.43)$$

Motivations:

- maximum of the entropy,
- central limit theorem.



Histogram of the mean of N uniform random variables

The gaussian distribution : Properties

- The law is a function of the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$:

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{2.44}$$

- The expectation of $\mathbf{x}$ under the Gaussian distribution is:

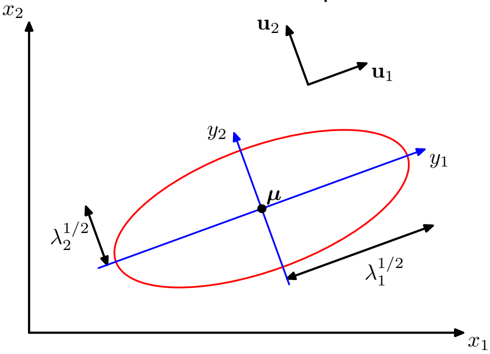
$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}, \tag{2.59}$$

- The covariance matrix of $\mathbf{x}$ is:

$$\text{cov}(\mathbf{x}) = \boldsymbol{\Sigma}. \tag{2.64}$$

The gaussian distribution : Properties

The law is constant on elliptical surfaces



where

- λ_i are the eigenvalues of $\mathbf{\Sigma}$,
- $\mathbf{u}_i$ are the associated eigenvectors.

The gaussian distribution : Conditional and marginal laws

Given a Gausian distribution $N(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with:

$$\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)^\top,\ \boldsymbol{\mu} = (\boldsymbol{\mu}_a, \boldsymbol{\mu}_b)^\top \tag{2.94}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{2.95}$$

- The conditional distribution $p(\mathbf{x}_a|\mathbf{x}_b)$ is a gaussian law with parameters:

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b), \tag{2.96}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \tag{2.82}$$

- The marginal distribution $p(\mathbf{x}_a)$ is a gaussian law with parameters $(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$.

The gaussian distribution : Bayes' theorem

A linear gaussian model is a couple of vectors $(\mathbf{x}, \mathbf{y})$ described by the relations:

$$p(\mathbf{x}) = N(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = N(\mathbf{y}, \mathbf{A}\mathbf{x} + \mathbf{b}, L^{-1}) \tag{2.114}$$

$(\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \boldsymbol{\epsilon})$ where $\mathbf{x}$ is gaussian and $\boldsymbol{\epsilon}$ is a centered gaussian noise).

Then

$$p(\mathbf{y}) = N(\mathbf{y}, \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top}) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = N(\mathbf{x}|\boldsymbol{\Sigma}(\mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}), \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1} \tag{2.117}$$

The gaussian distribution : Maximum likehood

Assume we have $\mathbf{X}$ a set of N iid observations following a Gaussian law. The parameters of the law, estimated by ML are:

$$\boldsymbol{\mu}_{\mathsf{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \tag{2.121}$$

$$\boldsymbol{\Sigma}_{\mathsf{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathsf{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathsf{ML}})^{\top}. \tag{2.122}$$

The empirical mean is unbiased but it is not the case of the empirical variance. The bias can be correct multiplying $\boldsymbol{\Sigma}_{\mathsf{ML}}$ by the factor $\frac{N}{N-1}$.

The gaussian distribution : Maximum likehood

The mean estimated form N data points is a revision of the estimator obtained from the $(N-1)$ first data points:

$$\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} = \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}). \qquad (2.126)$$

It is a particular case of the algorithm of Robbins-Monro, which iteratively search the root of a regression function.

The gaussian distribution : bayesian inference

- The conjugate prior for μ is gaussian,
- The conjugate prior for $\lambda = \frac{1}{\sigma^2}$ is a Gamma law,
- The conjugate prior of the couple (μ, λ) is the normal gamma distribution $N(\mu|\mu_0, \lambda_0^{-1})\text{Gam}(\lambda|a, b)$ where λ_0 is a linear function of λ.
- The posterior distribution would exhibit a coupling between the precision of μ and λ.
- The multidimensional conjugate prior is the Gaussian Wishart law.

The Gaussian distribution : limitations

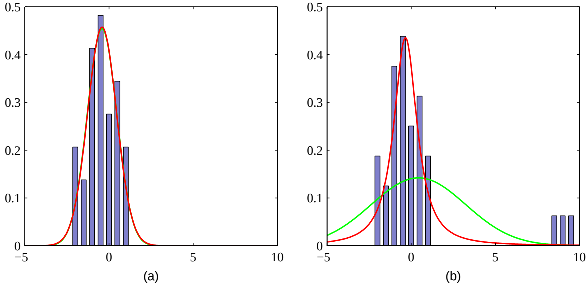
- A lot of parameters to estimate $D(1 + (D + 1)/2)$: simplification (diagonal variance matrix),
- Maximum likehood estimators are not robust to outliers: t-Student distribution,
- Not able to describe periodic data: von Mises distribution,
- Unimodal distribution Mixture of Gaussian.

After the gaussian distribution : t-Student distribution

- A student distribution is an infinite sum of gaussian having the same mean but different precisions (described by a Gamma law)

$$p(x|\mu, a, b) = \int_0^{\infty} N(x|\mu, \tau^{-1})\text{Gam}(\tau|a, b)d\tau \qquad (2.158)$$

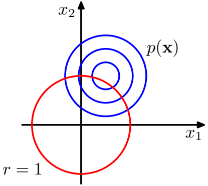
- It is robust to outliers



Histogram of 30 "gaussian" data points (+3 outliers) and ML estimator of the Gaussian (green) and the Student (red) laws

After the gaussian distribution : von Mises distribution

- When the data are periodic, it is necessary to work with polar coordinates.
- The von Mises law is obtained by conditionning the bidimensional gaussian law to the unit circle:



- the distribution is:

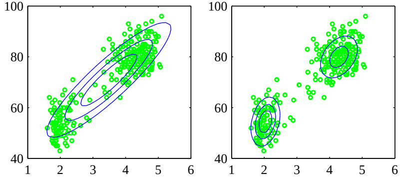
$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp(m \cos(\theta - \theta_0) \qquad (2.179)$$

where

  - m is the concentration (precision) parameter,
  - θ_0 is the mean.

Mixtures (of Gaussians) (1/3)

- Data with distinct regimes better modeled with mixtures



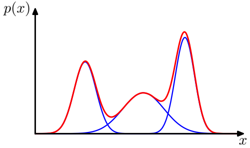
- General form: convex combination of component densities

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}), \qquad (2.188)$$

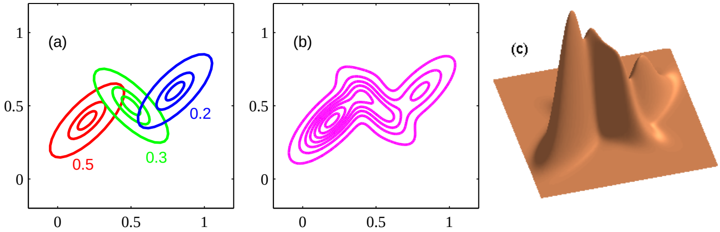
$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1, \quad \int p_k(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$$

Mixtures (of Gaussians) (2/3)

- Gaussian popular density, and so are mixtures thereof
- Example of mixture of Gaussians on $\mathbb{R}$



- Example of mixture of Gaussians on $\mathbb{R}^2$



Mixtures (of Gaussians) (3/3)

- Interpretation of mixture density: $p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$
  - mixing weight π_k is the prior probability $p(k)$ on the regimes
  - $p_k(\mathbf{x})$ is the conditional distribution $p(\mathbf{x}|k)$ on $\mathbf{x}$ given regime
  - $p(\mathbf{x})$ is the marginal on $\mathbf{x}$
  - $p(k|\mathbf{x}) \propto p(k)p(\mathbf{x}|k)$ is the posterior on the regime given $\mathbf{x}$

- The log-likelihood contains a log-sum

$$\log p(\{\mathbf{x}_n\}_{n=1}^{N}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_n) \qquad (2.193)$$

  - introduces local maxima and prevents closed-form solutions
  - iterative methods: gradient-ascent or bound-maximization
  - the posterior $p(k|\mathbf{x})$ appears in gradient and in (EM) bounds

The Exponential Family (1/3)

- Large family of useful distributions with common properties
  - Bernoulli, beta, binomial, chi-square, Dirichlet, gamma, Gaussian, geometric, multinomial, Poisson, Weibull, ...
  - Not in the family: Cauchy, Laplace, mixture of Gaussians, ...
  - Variable can be discrete or continuous (or vectors thereof)
- General form: log-linear interaction

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\} \tag{2.194}$$

- Normalization determines form of g:

$$g(\boldsymbol{\eta})^{-1} = \int h(\mathbf{x})\exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\}\,\mathrm{d}\mathbf{x} \tag{2.195}$$

  - Differentiation with respect to $\boldsymbol{\eta}$, using Leibniz's rule, reveals

$$-\nabla \log g(\boldsymbol{\eta}) = \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta})}\left[\mathbf{u}(\mathbf{x})\right] \tag{2.226}$$

The Exponential Family (2/3): Sufficient Statistics

- Maximum likelihood estimation for i.i.d. data $X = \{\mathbf{x}_n\}_{n=1}^N$

$$p(X) = \left(\prod_{n=1}^N h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^\top \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)\right\} \quad (2.227)$$

- Setting gradient w.r.t. $\boldsymbol{\eta}$ to zero yields

$$-\nabla \log g(\boldsymbol{\eta}_{ML}) = \frac{1}{N}\sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \quad (2.228)$$

  - $\sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$ is all we need from the data: sufficient statistics

- Combining with result from previous slide, ML estimate yields

$$\mathbb{E}_{p(\mathbf{x}|\boldsymbol{\eta}_{ML})}\left[\mathbf{u}(\mathbf{x})\right] = \frac{1}{N}\sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$$

The Exponential Family (3/3): Conjugate Priors

- Given a probability distribution $p(\mathbf{x}|\boldsymbol{\eta})$, prior $p(\boldsymbol{\eta})$ is conjugate if the posterior $p(\boldsymbol{\eta}|\mathbf{x})$ has the same form as the prior.
- All exponential family members have conjugate priors:

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu) g(\boldsymbol{\eta})^{\nu} \exp\left\{\nu\boldsymbol{\eta}^{\top}\boldsymbol{\chi}\right\} \qquad (2.229)$$

- Combining the prior with a exponential family likelihood

$$p(X=\{\mathbf{x}_n\}_{n=1}^{N}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^{N} \exp\left\{\boldsymbol{\eta}^{\top}\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right\}$$

we obtain (2.230)

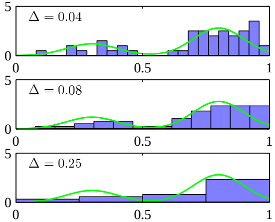
$$p(\boldsymbol{\eta}|X,\boldsymbol{\chi},\nu) \propto g(\boldsymbol{\eta})^{N+\nu} \exp\left\{\boldsymbol{\eta}^{\top}\left(\nu\boldsymbol{\chi}+\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)\right)\right\}$$

Nonparametric methods

- So far we have seen parametric densities in this chapter
  - Limitation: we are tied down to a specific functional form
  - Alternatively we can use (flexible) nonparametric methods
- Basic idea: consider small region $\mathcal{R}$, with $P = \int_{\mathcal{R}} p(\mathbf{x})\, \mathrm{d}\mathbf{x}$
  - For $N \to \infty$ data points we find about $K \approx NP$ in $\mathcal{R}$
  - For small $\mathcal{R}$ with volume V: $P \approx p(\mathbf{x})V$ for $\mathbf{x} \in \mathcal{R}$
  - Thus, combining we find: $p(\mathbf{x}) \approx K/(NV)$
- Simplest example: histograms
  - Choose bins
  - Estimate density in i-th bin

$$p_i = \frac{n_i}{N\Delta_i} \qquad (2.241)$$

  - Tough in many dimensions: smart chopping required



Kernel density estimators: fix V, find K

- Let $\mathcal{R} \in \mathbb{R}^D$ be a unit hypercube around $\mathbf{x}$, with indicator

$$k(\mathbf{x}-\mathbf{y}) = \begin{cases} 1 & : \quad |x_i - y_i| \leq 1/2 \quad (i = 1, \ldots, D) \\ 0 & : \quad \text{otherwise} \end{cases} \tag{2.247}$$

- $\#$ points in $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in hypercube of side h is:

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \tag{2.248}$$

- Plug this into approximation $p(\mathbf{x}) \approx K/(NV)$, with $V = h^D$:

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \tag{2.249}$$

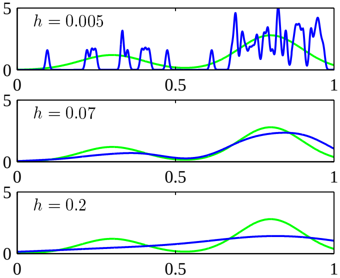
  - Note: this is a mixture density!

Kernel density estimators

- Smooth kernel density estimates obtained with Gaussian

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}\exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\} \quad (2.250)$$

- Example with Gaussian kernel for different values of the smoothing parameter h

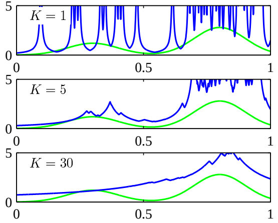


Nearest-neighbor methods: fix K, find V

- Single smoothing parameter for kernel approach is limiting
  - too large: structure is lost in high-density areas
  - too small: noisy estimates in low-density areas
  - we want density-dependent smoothing
- Nearest Neighbor method also based on local approximation:

$$p(\mathbf{x}) \approx K/(NV) \tag{2.246}$$

- For new $\mathbf{x}$, find the volume of the smallest circle centered on $\mathbf{x}$ enclosing K points

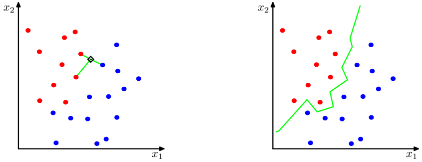


Nearest-neighbor methods: classification with Bayes rule

- Density estimates from K-neighborhood with volume V:
  - Marginal density estimate $p(\mathbf{x}) = K/(NV)$
  - Class prior esimates: $p(\mathcal{C}_k) = N_k/N$
  - Class-conditional estimate $p(\mathbf{x}|\mathcal{C}_k) = K_k/(N_k V)$
- Posterior class probability from Bayes rule:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathcal{C}_k)p(\mathbf{x}|\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K} \qquad (2.256)$$

  - Classification based on class-counts in K-neighborhood
  - In limit $N \to \infty$ classification error at most $2\times$ optimal [Cover & Hart, 1967]
- Example for binary classification, (a) $K = 3$, (b) $K = 1$



(a) (b)