Probability Distributions

Density estimation

To model the probability distribution of a random variable, given a finite set of observations.

Parametric approach — Specific functional form for the distribution.

Nonparametric approach — The form of the distribution typically depends on the size of the data set.

Frequentist: choose specific values for the parameters by optimizing some criterion, such as the likelihood function.
Bayesian: introduce prior distributions over the parameters and then use Bayes' theorem to compute the corresponding posterior distribution given the observed data.

Conjugate Prior: leads to a posterior distribution of the same functional form as the prior

Chapter 2 Probability Distributions

2.1 Binary Variables

2.2 Multinomial Variables

2.3 The Gaussian Distribution

2.4 The Exponential Family

2.5 Nonparametric Methods
- Kernel density estimators
- Nearest-neighbour methods

2.1 Binary Variables

A binary random variable $x \in \{0,1\}$, $p(x = 1|\mu) = \mu, p(x = 0|\mu) = 1 - \mu$

The probability distribution over x $\longrightarrow$ $\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$

Mean and variance:

$$\begin{aligned} \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu). \end{aligned}$$

Maximum likelihood estimator:
(Frequentist way)

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

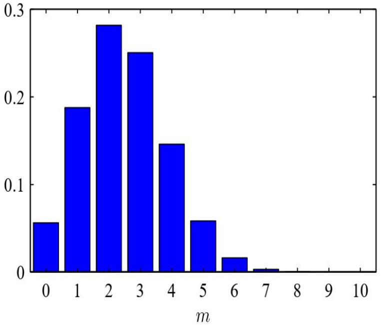
This can give severely over-fitted results for small data sets

2.1 Binary Variables

A binary random variable $x \in \{0,1\}$, $p(x=1|\mu)=\mu, p(x=0|\mu)=1-\mu$

The probability distribution over x $\longrightarrow$ $\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x}$

The distribution of the number m of observations of $x=1$: $\mathrm{Bin}(m|N,\mu)=\binom{N}{m}\mu^m(1-\mu)^{N-m}$



$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) \quad = \quad N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m-\mathbb{E}[m])^2\,\mathrm{Bin}(m|N,\mu) \quad = \quad N\mu(1-\mu)$$

Gamma function

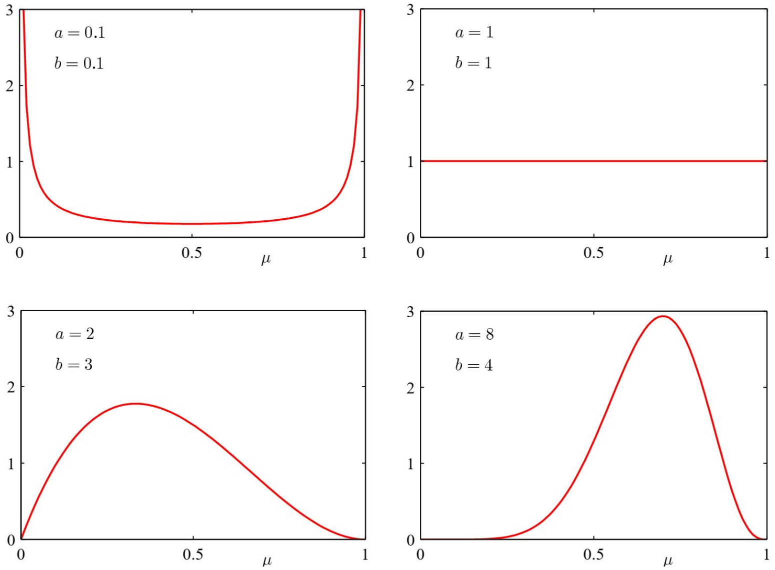
$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} \mathrm{d}u$$

$$\Gamma(x+1) = x\Gamma(x)$$

N is interger:

$$\Gamma(n) = (n-1)!$$

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



Chapter 2 Probability Distributions

2.1 Binary Variables

A binary random variable $x \in \{0,1\}$, $p(x=1|\mu)=\mu, p(x=0|\mu)=1-\mu$

The probability distribution over x $\longrightarrow$ $\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x}$

From the Bayesian perspective, we need to introduce a prior distribution $p(\mu)$ over the parameter μ.

Prior

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

$$\begin{aligned}\mathbb{E}[\mu] &= \frac{a}{a+b} \\ \mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}\end{aligned}$$

Posterior

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

- 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

https://en.wikipedia.org/wiki/Conjugate_prior

Chapter 2 Probability Distributions

2.2 Multinomial Variables

$$\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}} \qquad \sum_{k=1}^{K} x_k = 1 \qquad p(x_k = 1) = \mu_k$$

The distribution of x is given: $$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ $$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}.$$

The joint distribution of the quantities $m_1, \ldots, m_K$

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

Chapter 2 Probability Distributions

2.2 Multinomial Variables

Introduce a family of prior distributions for the parameters $\{\mu_k\}$ of the multinomial distribution.

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

Dirichlet distribution

Multiplying the prior by the likelihood function (2.34), we obtain the posterior distribution:

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

Dirichlet is indeed a conjugate prior for the multinomial.

Chapter 2 Probability Distributions

2.3 The Gaussian Distribution

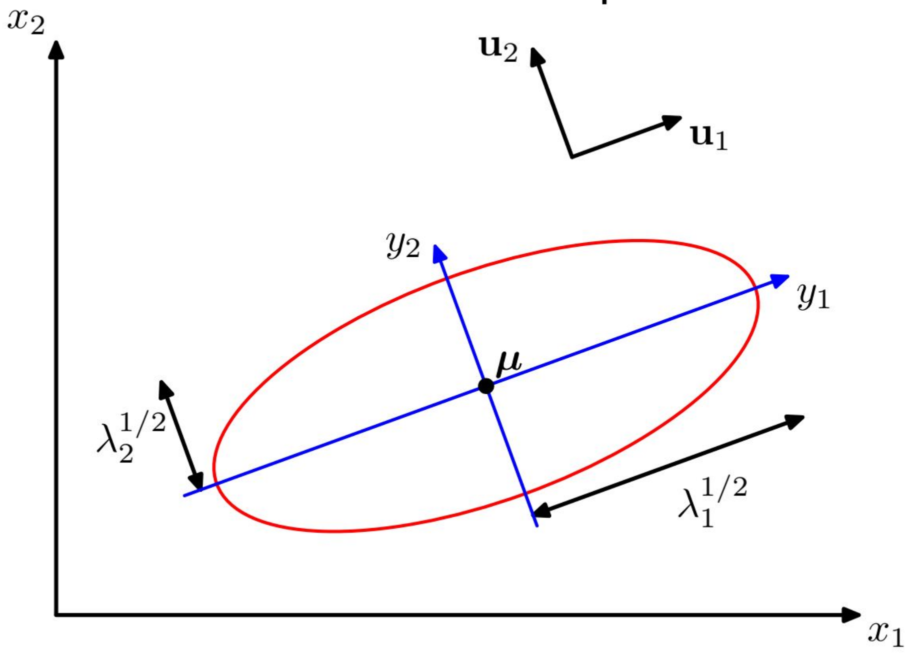
The multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Limitations:

- The total number of parameters grows quadratically with the dimension D.
- It is intrinsically unimodal.

The law is constant on elliptical surfaces



where

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

2.3.1 Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Conditional distribution:

$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b). \end{aligned}$$

A linear function of $\mathbf{x}_b$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}).$$

Chapter 2 Probability Distributions

2.3.3 Bayes' theorem for Gaussian variables

Given a Gaussian marginal distribution p(x) and a Gaussian conditional distribution $p(\mathbf{y}|\mathbf{x})$ which has a mean that is a linear function of x, and a covariance which is independent of x.

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

The evaluation of this conditional can be seen as an example of Bayes' theorem.

$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})
\end{aligned}
$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}.$$

We can interpret the distribution $p(\mathbf{x})$ as a prior distribution over $\mathbf{x}$.

2.3.4 Maximum likelihood for the Gaussian

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$

The observ $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma} \qquad \widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

Chapter 2 Probability Distributions

2.3.5 Sequential estimation

Sequential methods allow data points to be processed one at a time and then discarded and are important for on-line applications, and also where large data sets are involved so that batch processing of all data points at once is infeasible.

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)})
\end{aligned}
$$

However, we will not always be able to derive a sequential algorithm by this route, and so we seek a more general formulation of sequential learning.

Robbins-Monro procedure

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}\ln p(x_N|\theta^{(N-1)})$$

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.

Chapter 2 Probability Distributions

2.3.6 Bayesian inference for the Gaussian

The variance σ^2 is known, the mean μ is unknown.

Prior $$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right)$$

Posterior $$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

A sequential update formula:

$$p(\boldsymbol{\mu}|D) \propto \left[p(\boldsymbol{\mu}) \prod_{n=1}^{N-1} p(\mathbf{x}_n|\boldsymbol{\mu})\right] p(\mathbf{x}_N|\boldsymbol{\mu})$$

The variance σ^2 is unknown, the mean μ is known.

$$\lambda \equiv 1/\sigma^2 \qquad \mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$p(\lambda|\mathbf{x}) \propto \lambda^{a_0-1}\lambda^{N/2} \exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

$$\mathrm{Gam}(\lambda|a_N, b_N)$$

$$a_N = a_0 + \frac{N}{2}$$

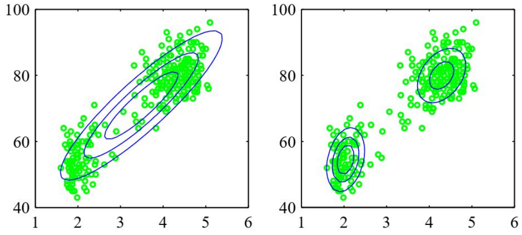
$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma_{\mathrm{ML}}^2$$

Both the mean and the precision are unknown,
normal-gamma

Chapter 2 Probability Distributions

2.3.9 Mixtures of Gaussians

- Data with distinct regimes better modeled with mixtures



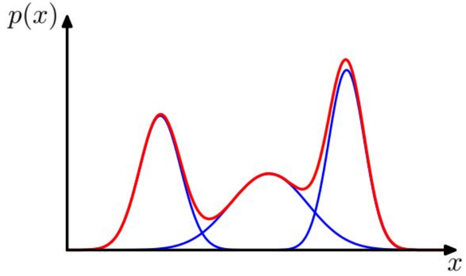
- General form: convex combination of component densities

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

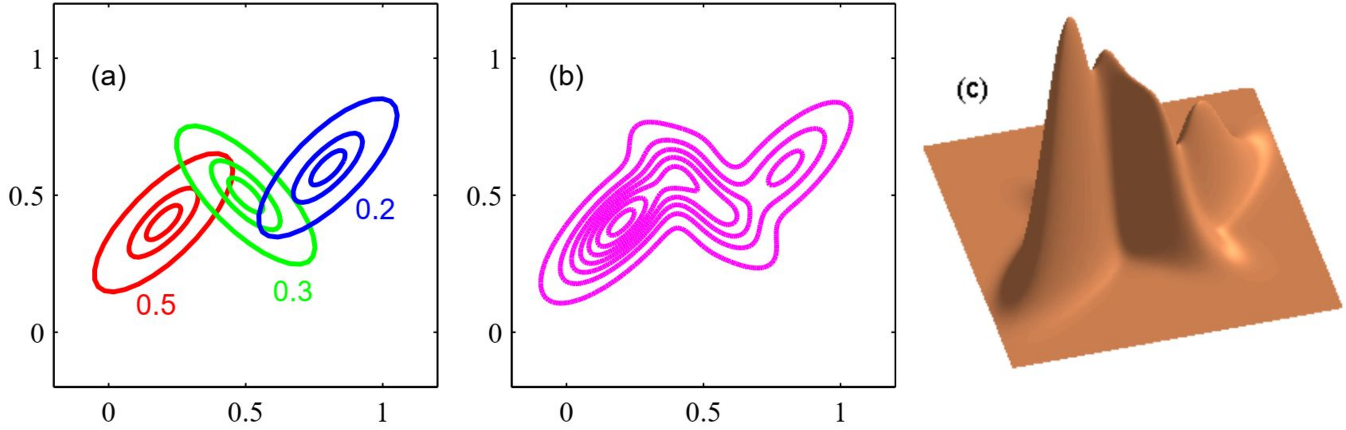
$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1, \quad \int p_k(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1$$

- Gaussian popular density, and so are mixtures thereof

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



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



- 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

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

  - 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

Chapter 2 Probability Distributions

2.4. The Exponential Family

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

Bernoulli distribution

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= \sigma(-\eta) \end{aligned}$$

Gaussian distribution

$$\begin{aligned} \boldsymbol{\eta} &= \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \\ \mathbf{u}(x) &= \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ h(\mathbf{x}) &= (2\pi)^{-1/2} \\ g(\boldsymbol{\eta}) &= (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \end{aligned}$$

multinomial distribution

$$\begin{aligned} \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \end{aligned}$$

Chapter 2 Probability Distributions

2.4.1 Maximum likelihood and sufficient statistics

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

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

Setting the gradient of ln $p(\mathbf{X}|\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ to zero

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

The only we need. ***sufficient statistic*** of exponential family.

Chapter 2 Probability Distributions

2.4.2 Conjugate priors

Given a probability distribution $\boldsymbol{p(x|\eta)}$, if the prior $\boldsymbol{p(\eta)}$ is conjugate, the posterior has the same form as the prior.

All exponential family members have conjugate priors:

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

Combining the prior with a exponential family likelihoo

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

we obtain

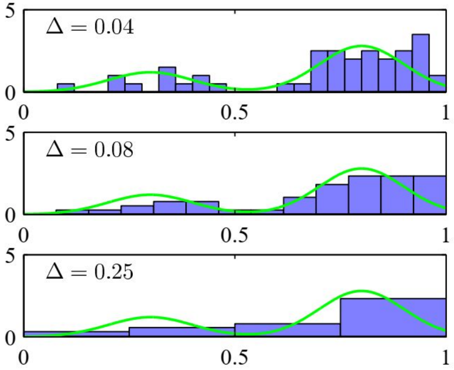
$$p(\boldsymbol{\eta}|\mathbf{X},\boldsymbol{\chi},\nu) \propto g(\boldsymbol{\eta})^{\nu+N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\left(\sum_{n=1}^{N}\mathbf{u}(\mathbf{x}_n)+\nu\boldsymbol{\chi}\right)\right\}$$

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 $V = h^D$:

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

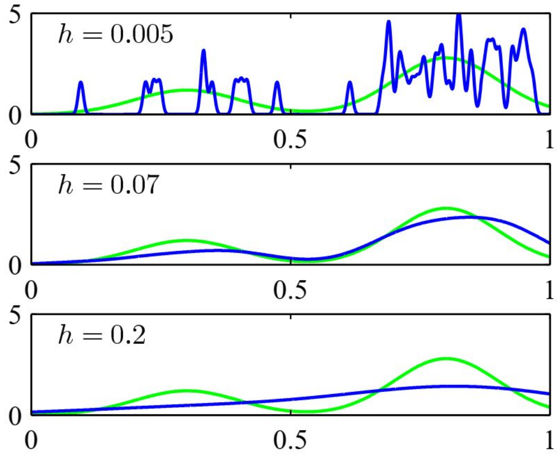
  - Note: this is a mixture density!

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

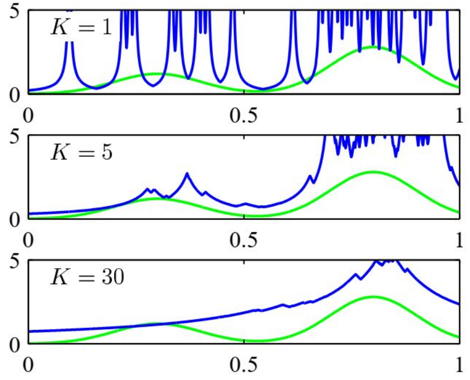


Nearest-neighbor methods: fix K, find V

- 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) $K = 1$