MACHINE LEARNING

CHAPTER 2: PROBABILITY DISTRIBUTIONS

Learning Objectives

1、 What are binary, multinomial and Gaussian distributions and their conjugate prior distributions?

2、What are the common properties of Gaussian distributions?

3、What are exponential families and their properties?

4、How to choose non-informative prior*?

5、How to use non-parametric methods for learning?

6、What are KNN based methods?

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

The Exponential Family (1)

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

where $\boldsymbol{\eta}$ is the *natural parameter* and

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

so $\mathrm{g}(\boldsymbol{\eta})$ can be interpreted as a normalization coefficient.

$\mathbf{u}(\mathbf{x})$: statistics of $\mathbf{x}$

The Exponential Family (2.1)

The Bernoulli Distribution

$$
\begin{aligned}
p(x|\mu) &= \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \\
&= \exp\left\{x \ln \mu + (1-x)\ln(1-\mu)\right\} \\
&= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}
\end{aligned}
$$

Comparing with the general form we see that

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$$ and so $$\mu = \underbrace{\sigma(\eta) = \frac{1}{1+\exp(-\eta)}}_{\text{Logistic sigmoid}}.$$

The Exponential Family (2.2)

The Bernoulli distribution can hence be written as

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

where

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

The Exponential Family (3.1)

The Multinomial Distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right)$$

where, $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$ and

$$\begin{aligned} \eta_k &= \ln \mu_k \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= 1. \end{aligned}$$

NOTE: The η_k parameters are not independent since the corresponding μ_k must satisfy

$$\sum_{k=1}^{M} \mu_k = 1.$$

The Exponential Family (3.2)

Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$. This leads to

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \quad \text{and} \quad \mu_k = \underbrace{\frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}}_{\text{Softmax}}.$$

Here the η_k parameters are independent. Note that

$$0 \leqslant \mu_k \leqslant 1 \quad \text{and} \quad \sum_{k=1}^{M-1} \mu_k \leqslant 1.$$

The Exponential Family (3.3)

The Multinomial distribution can then be written as

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

where

$$\begin{aligned}
\boldsymbol{\eta} &= (\eta_1, \ldots, \eta_{M-1}, 0)^{\mathrm{T}} \\
\mathbf{u}(\mathbf{x}) &= \mathbf{x} \\
h(\mathbf{x}) &= 1 \\
g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1}.
\end{aligned}$$

The Exponential Family (4)

The Gaussian Distribution

$$\begin{aligned} p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \\ &= h(x)g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(x)\right\} \end{aligned}$$

where

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \qquad h(\mathbf{x}) = (2\pi)^{-1/2}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right).$$

ML for the Exponential Family (1)*

From the definition of $\mathrm{g}(\boldsymbol{\eta})$ we get

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x}}_{1/g(\boldsymbol{\eta})} + \underbrace{g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

ML for the Exponential Family (2)*

Give a data set, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the likelihood function is given by

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

Thus we have

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

Conjugate priors

For any member of the exponential family, there exists a prior

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

Combining with the likelihood function, we get

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

Prior corresponds to ν pseudo-observations with value χ.

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

Non-informative Priors (1)*

With little or no information available a-priori, we might choose a non-informative prior.

- λ discrete, K-nomial : $p(\lambda) = 1/K$.
- $\lambda \in [a,b]$ real and bounded: $p(\lambda) = 1/b - a$.
- λ real and unbounded: improper!

A constant prior may no longer be constant after a change of variable; consider $p(\lambda)$ constant and $\lambda = \eta^2$:

$$p_\eta(\eta) = p_\lambda(\lambda)\left|\frac{\mathrm{d}\lambda}{\mathrm{d}\eta}\right| = p_\lambda(\eta^2)2\eta \propto \eta$$

Non-informative Priors (2)*

Translation invariant priors. Consider

$$p(x|\mu) = f(x - \mu) = f((x + c) - (\mu + c)) = f(\widehat{x} - \widehat{\mu}) = p(\widehat{x}|\widehat{\mu}).$$

For a corresponding prior over μ, we have

$$\int_A^B p(\mu)\,\mathrm{d}\mu = \int_{A-c}^{B-c} p(\mu)\,\mathrm{d}\mu = \int_A^B p(\mu - c)\,\mathrm{d}\mu$$

for any A and B. Thus $p(\mu) = p(\mu - c)$ and $p(\mu)$ must be constant.

Non-informative Priors (3)*

Example: The mean of a Gaussian, μ; the conjugate prior is also a Gaussian,

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

As $\sigma_0^2 \to \infty$, this will become constant over μ.

Non-informative Priors (4)*

Scale invariant priors. Consider $p(x|\sigma) = (1/\sigma) f(x/\sigma)$ and make the change of variable $\widehat{x} = cx$

$$p_{\widehat{x}}(\widehat{x}) = p_x(x) \left| \frac{\mathrm{d}x}{\mathrm{d}\widehat{x}} \right| = p_x\left(\frac{\widehat{x}}{c}\right)\frac{1}{c} = \frac{1}{c\sigma} f\left(\frac{\widehat{x}}{c\sigma}\right) = p_x(\widehat{x}|\widehat{\sigma}).$$

For a corresponding prior over σ, we have

$$\int_A^B p(\sigma)\,\mathrm{d}\sigma = \int_{A/c}^{B/c} p(\sigma)\,\mathrm{d}\sigma = \int_A^B p\left(\frac{1}{c}\sigma\right)\frac{1}{c}\,\mathrm{d}\sigma$$

for any A and B. Thus $p(\sigma) \; / \; 1/\sigma$ and so this prior is improper too. Note that this corresponds to $p(\ln\sigma)$ being constant.

Non-informative Priors (5)*

Example: For the variance of a Gaussian, σ^2, we have

$$\mathcal{N}(x|\mu,\sigma^2) \propto \sigma^{-1} \exp\left\{-((x-\mu)/\sigma)^2\right\}.$$

If $\lambda = 1/\sigma^2$ and $p(\sigma) \;/\; 1/\sigma$, then $p(\lambda) \;/\; 1/\lambda$.

- We know that the conjugate distribution for λ is the Gamma distribution,

$$\mathrm{Gam}(\lambda|a_0,b_0) \propto \lambda^{a_0-1}\exp(-b_0\lambda).$$

- A non-informative prior is obtained when $a_0 = 0$ and $b_0 = 0$.

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-information Priors
- Non-parametric Methods
- KNN

Non-parametric Methods (1)

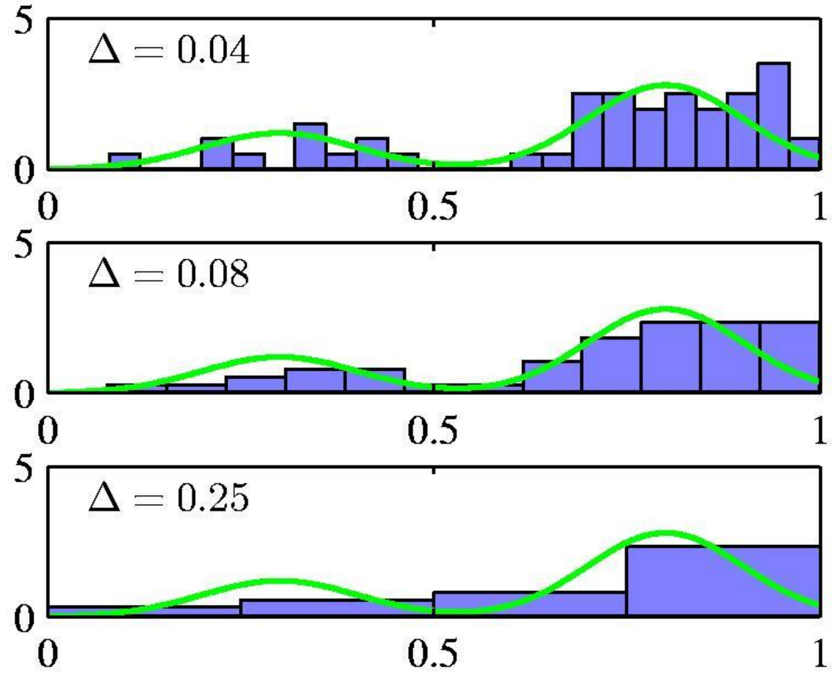
- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Non-parametric approaches make few assumptions about the overall shape of the distribution being modelled.

Non-parametric Methods (2)

Histogram methods partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

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

- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.



- In a D-dimensional space, using M bins in each dimension will require M^D bins!

Non-parametric Methods (3)

- Assume observations drawn from a density $p(\mathrm{x})$ and consider a small region R containing $\mathbf{x}$ such that

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

- The probability that K out of N observations lie inside R is $\mathrm{Bin}(K|N,P)$ and if N is large

$$K \simeq NP.$$

- If the volume of R, V, is sufficiently small, $p(\mathbf{x})$ is approximately constant over R and

$$P \simeq p(\mathbf{x})V$$

Thus

$$\boxed{p(\mathbf{x}) = \frac{K}{NV}.}$$

V small, yet $K>0$, therefore N large?

Non-parametric Methods (4)

Kernel Density Estimation: fix V, estimate K from the data. Let $\mathcal{R}$ be a hypercube centred on $\mathbf{x}$ and define the kernel function (Parzen window)

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$ and hence $$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$

Non-parametric Methods (5)

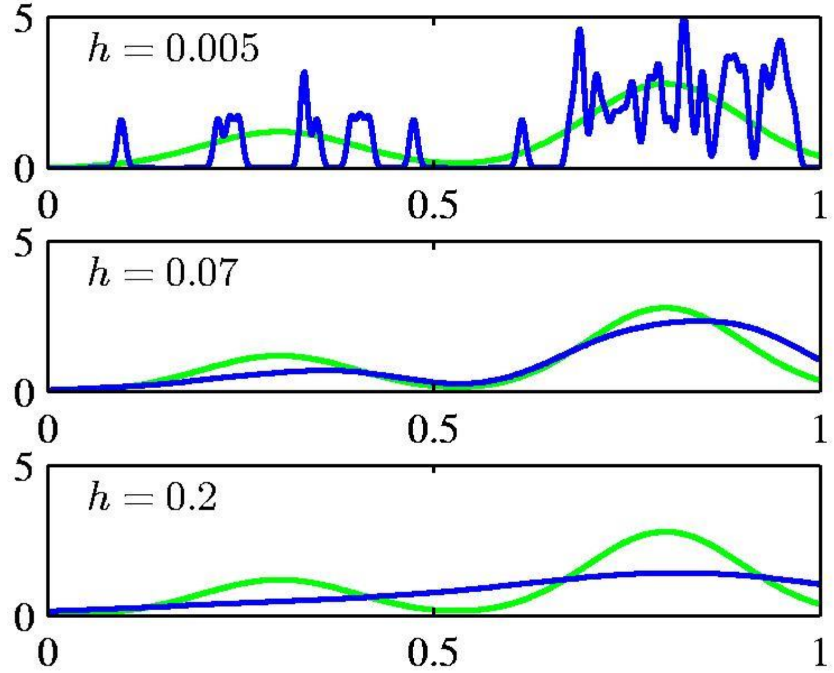
To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

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

Any kernel such that

$$\begin{aligned} k(\mathbf{u}) &\geqslant 0, \\ \int k(\mathbf{u})\,\mathrm{d}\mathbf{u} &= 1 \end{aligned}$$

will work.

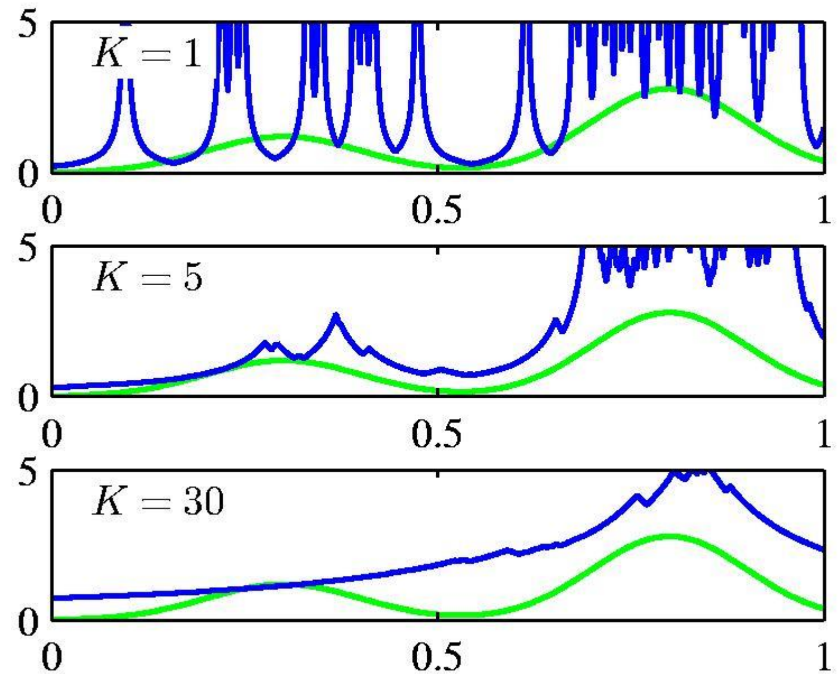


h acts as a smoother.

Non-parametric Methods (6)

Nearest Neighbour Density Estimation: fix K, estimate V from the data. Consider a hypersphere centred on x and let it grow to a volume, V^*, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$



K acts as a smoother.

Non-parametric Methods (7)

- Nonparametric models (not histograms) requires storing and computing with the entire data set.
- Parametric models, once fitted, are much more efficient in terms of storage and computation.

Outlines

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-informative Priors
- Non-parametric Methods
- KNN

K-Nearest-Neighbours for Classification (1)

- Given a data set with N_k data points from class $\mathrm{C_k}$, we have $\sum_k N_k = N$

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

number of data in a region

number of total data

volume of the region

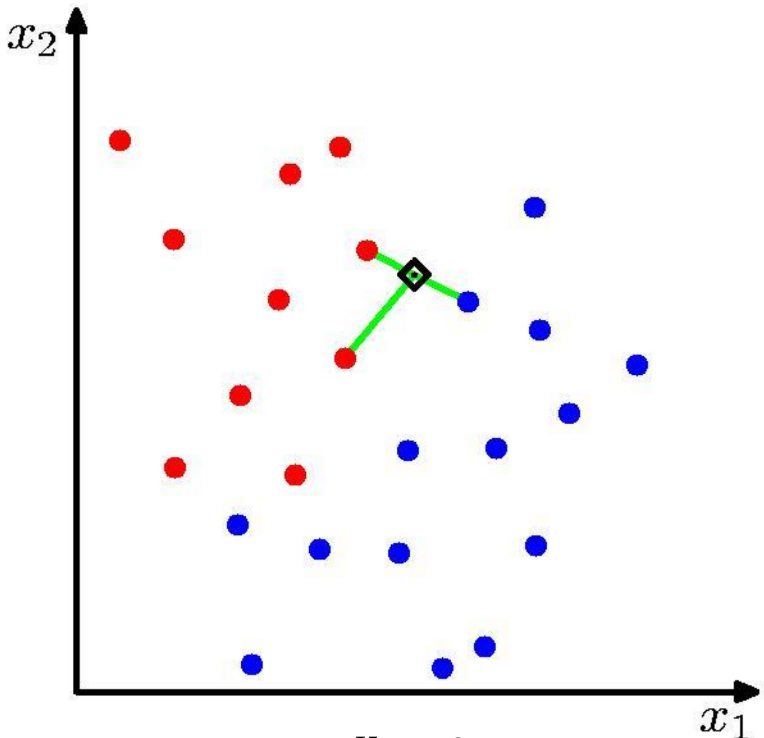
and correspondingly

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

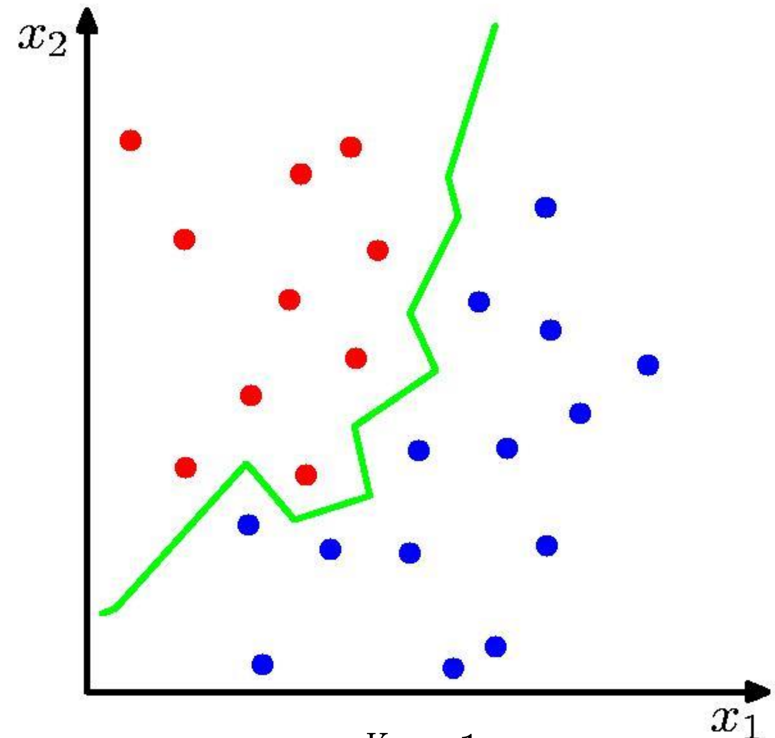
- Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

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

K-Nearest-Neighbours for Classification (2)

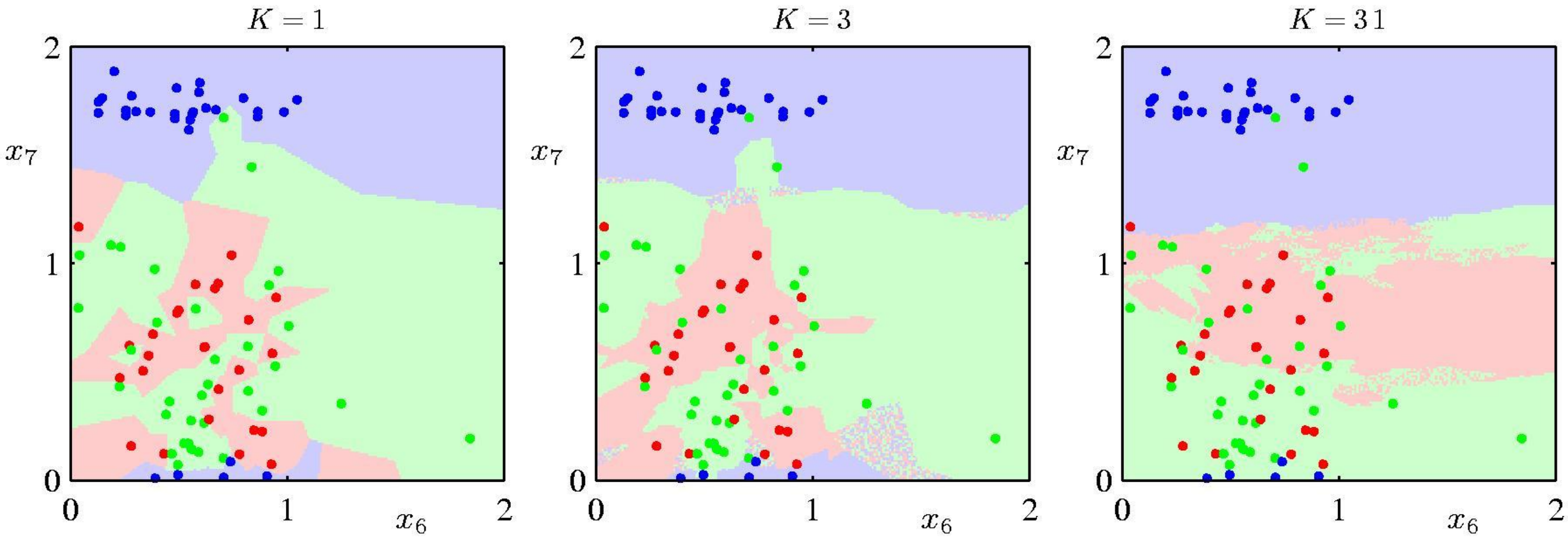


$K = 3$



$K = 1$

K-Nearest-Neighbours for Classification (3)



• K acts as a smother

• For $N \to \infty$, the error rate of the 1-nearest-neighbour classifier is never more than twice the optimal error (obtained from the true conditional class distributions).

Summary

- Binary Distributions
- Multinomial Distributions
- Gaussian Distributions
- Exponential Families
- Non-information Priors
- Non-parametric Methods
- KNN