Unsupervised Learning and Probability Review

Neil D. Lawrence

Department of Computer Science, University of Sheffield, U.K.

25th October 2013

Outline

Basic Probability

Basic Probability

Probability Density Functions

Clustering

- Divide data into discrete groups according to characteristics.
  - For example different animal species.
  - Different political parties.
- Determine the allocation to the groups and (harder) number of different groups.

K-means Clustering
An Algorithm

- *Require*: Set of K cluster centers & assignment of each point to a cluster.
  - Initialize cluster centers as data points.
  - Assign each data point to nearest cluster center.
  - Update each cluster center by setting it to the mean of assigned data points.

Objective Function

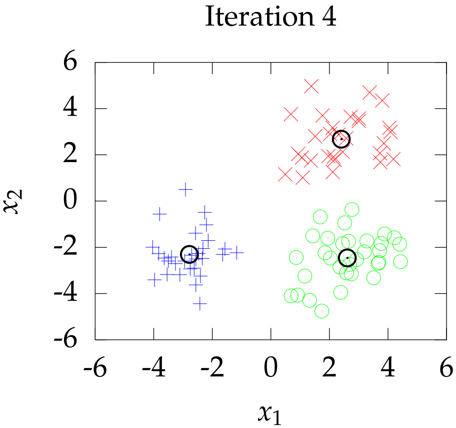
- This minimizes the objective:

$$\sum_{j=1}^{K} \sum_{i \text{ allocated to } j} \left(\mathbf{x}_{i,:} - \boldsymbol{\mu}_{j,:}\right)^{\top} \left(\mathbf{x}_{i,:} - \boldsymbol{\mu}_{j,:}\right)$$

- i.e. it minimizes the sum of Euclidean squared distances between points and their associated centers.
- The minimum is not guaranteed to be *global* or *unique*.
  - This objective is a non-convex optimization problem.

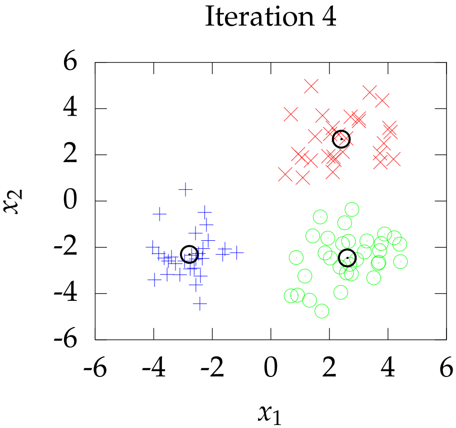
K-means Clustering

- K-means clustering.
  - Update each center by setting to the mean of the allocated points.



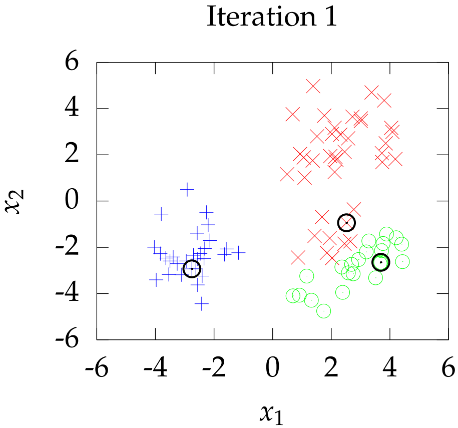
K-means Clustering

- K-means clustering.
  - Allocate each data point to the nearest cluster center.



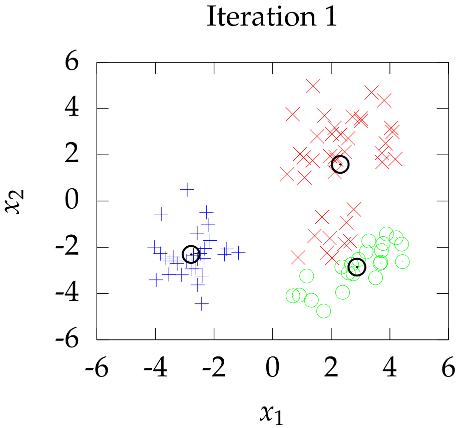
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



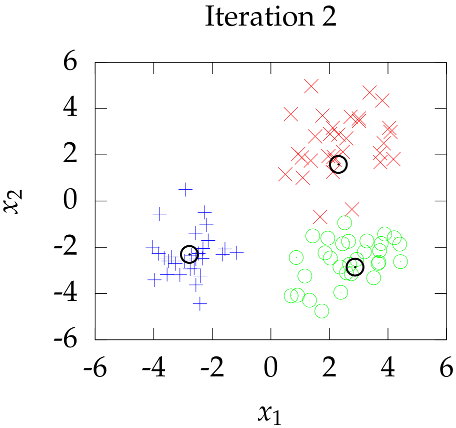
K-means Clustering

- K-means clustering.
  - Allocate each data point to the nearest cluster center.



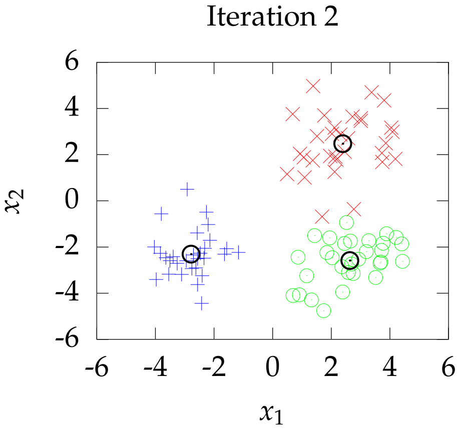
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



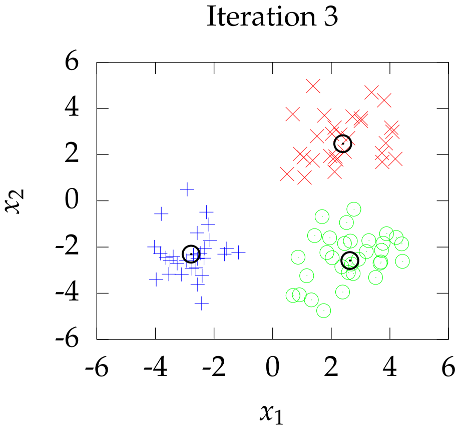
K-means Clustering

- K-means clustering.
  - Allocate each data point to the nearest cluster center.



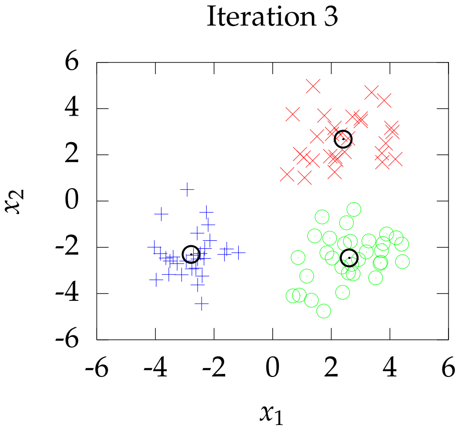
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



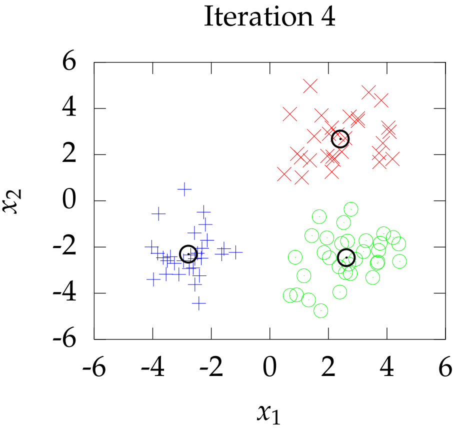
K-means Clustering

- K-means clustering.
  - Allocate each data point to the nearest cluster center.



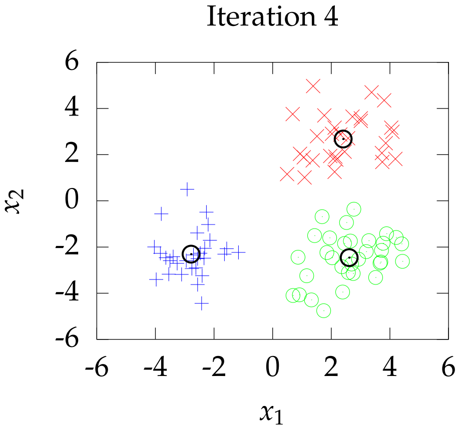
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



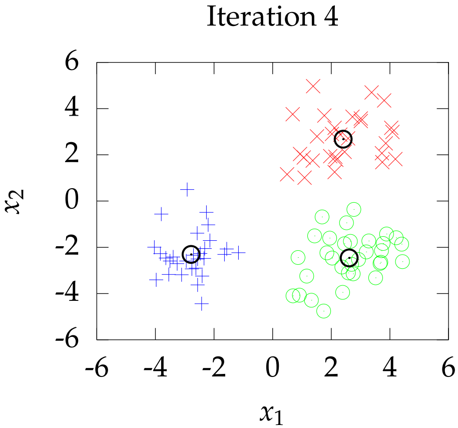
K-means Clustering

- *K*-means clustering.
  - Allocate each data point to the nearest cluster center.



K-means Clustering

- *K*-means clustering.
  - Allocation doesn't change so stop.



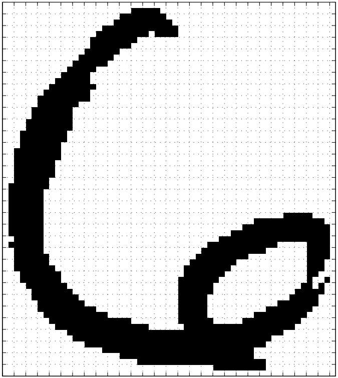
Other Clustering Approaches

- Spectral clustering (Shi and Malik, 2000; Ng et al., 2002).
  - Allows clusters which aren't convex hulls.
- Dirichlet processes
  - A probabilistic formulation for a clustering algorithm that is non-parameteric.

High Dimensional Data

USPS Data Set Handwritten Digit

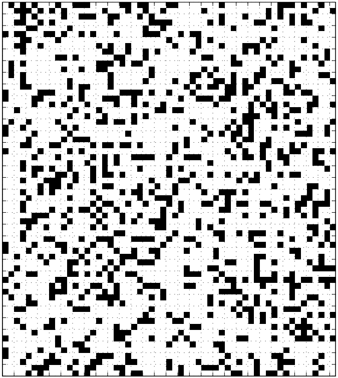
- 3648 Dimensions
- 64 rows by 57 columns



High Dimensional Data

USPS Data Set Handwritten Digit

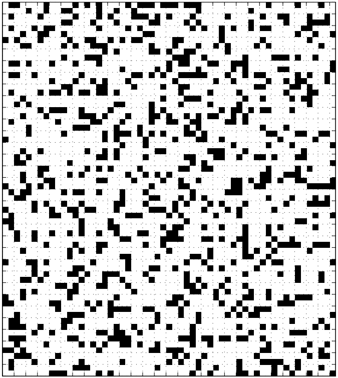
- 3648 Dimensions
- 64 rows by 57 columns
- Space contains more than just this digit.



High Dimensional Data

USPS Data Set Handwritten Digit

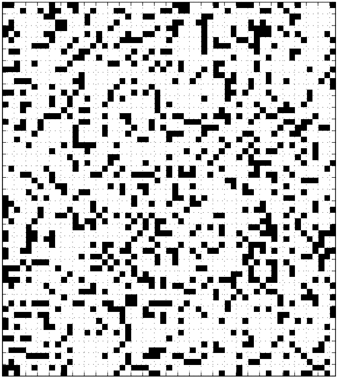
- 3648 Dimensions
- 64 rows by 57 columns
- Space contains more than just this digit.
- Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



High Dimensional Data

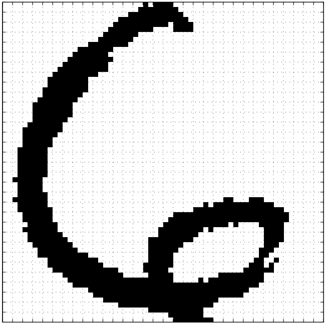
USPS Data Set Handwritten Digit

- 3648 Dimensions
- 64 rows by 57 columns
- Space contains more than just this digit.
- Even if we sample every nanosecond from now until the end of the universe, you won't see the original six!



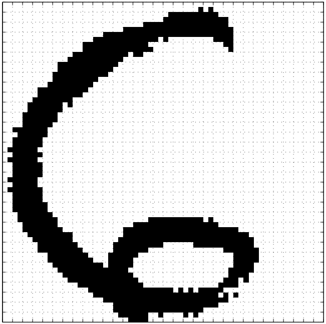
Simple Model of Digit

- Rotate a 'Prototype'



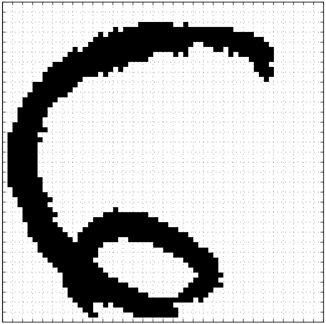
Simple Model of Digit

- Rotate a 'Prototype'



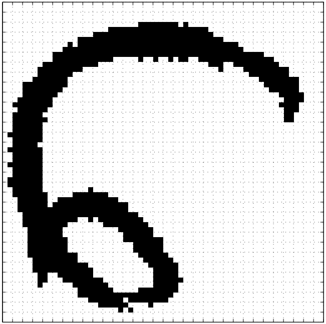
Simple Model of Digit

- Rotate a 'Prototype'



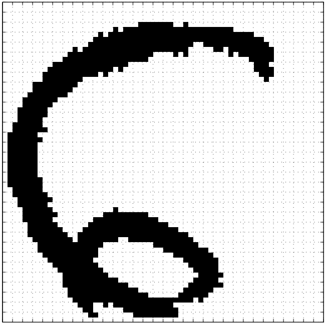
Simple Model of Digit

- Rotate a 'Prototype'



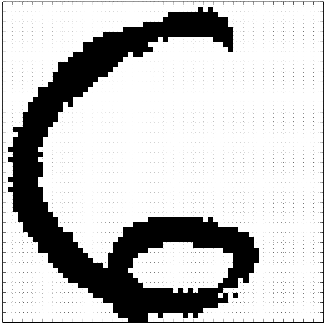
Simple Model of Digit

- Rotate a 'Prototype'



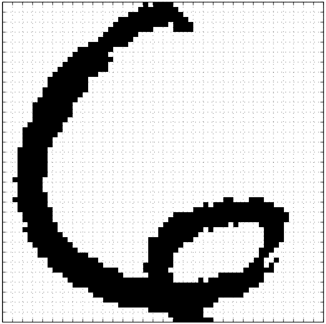
Simple Model of Digit

- Rotate a 'Prototype'



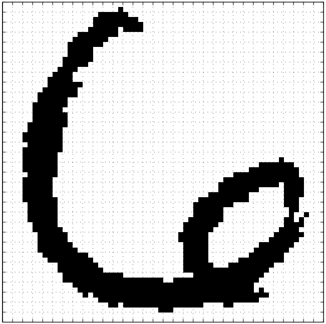
Simple Model of Digit

- Rotate a 'Prototype'



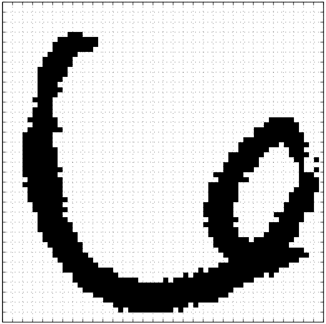
Simple Model of Digit

- Rotate a 'Prototype'



Simple Model of Digit

- Rotate a 'Prototype'

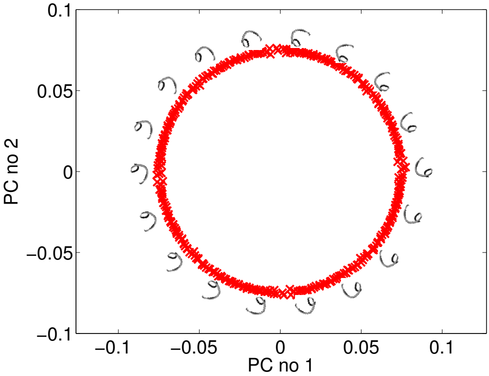


MATLAB Demo

```
demDigitsManifold([1 2], 'all')
```

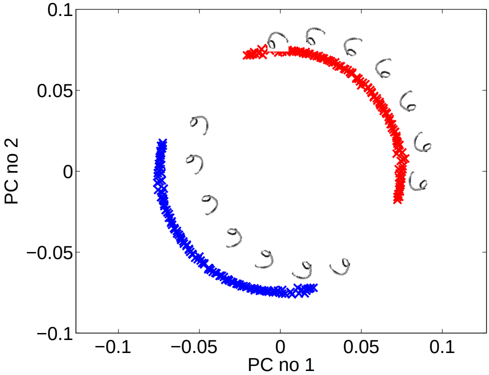
MATLAB Demo

```
demDigitsManifold([1 2], 'all')
```



MATLAB Demo

```
demDigitsManifold([1 2], 'sixnine')
```



Low Dimensional Manifolds

Pure Rotation is too Simple

- In practice the data may undergo several distortions.
  - *e.g.* digits undergo 'thinning', translation and rotation.
- For data with 'structure':
- we expect fewer distortions than dimensions;
- we therefore expect the data to live on a lower dimensional manifold.
- Conclusion: deal with high dimensional data by looking for lower dimensional non-linear embedding.

Principal Component Analysis

- How do we find these directions?
- Rotate to find directions in data with maximal variance.
  - This is known as PCA (Hotelling, 1933).
- Rotate data to extract directions of maximum variance.
- Do this by diagonalizing the sample covariance matrix

$$\mathbf{S} = n^{-1} \sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^\top$$

Principal Component Analysis

- Find a direction in the data, $\mathbf{x} = \mathbf{R}\mathbf{x}$, for which variance is maximized.

Lagrangian

- Solution is found via constrained optimisation (which uses *Lagrange* multipliers):

$$L\left(\mathbf{r}_1, \lambda_1\right) = \mathbf{r}_1^\top \mathbf{S}\mathbf{r}_1 + \lambda_1 \left(1 - \mathbf{r}_1^\top \mathbf{r}_1\right)$$

- Gradient with respect to $\mathbf{r}_1$

$$\frac{\mathrm{d}L\left(\mathbf{r}_1, \lambda_1\right)}{\mathrm{d}\mathbf{r}_1} = 2\mathbf{S}\mathbf{r}_1 - 2\lambda_1\mathbf{r}_1$$

  rearrange to form

$$\mathbf{S}\mathbf{r}_1 = \lambda_1\mathbf{r}_1.$$

  Which is known as an *eigenvalue* problem.

- Further directions can also be shown to be eigenvectors of the covariance.

Error Functions to Probabilities

- We introduced different learning scenarios using error functions.
- Now we will reinterpret those error functions through probability.
- The error function can be seen as a logarithm of a probability density function.
- Before we take that perspective we will first review probability.

Outline

Basic Probability

Basic Probability

Probability Density Functions

Probability Review I

- We are interested in trials which result in two random variables, Y and X, each of which has an 'outcome' denoted by y or x.
- We summarise the notation and terminology for these distributions in the following table.

Terminology	Notation	Description
Joint Probability	$P(Y = y, X = x)$	'The probability that $Y = y$ *and* $X = x$'
Marginal Probability	$P(Y = y)$	'The probability that $Y = y$ *regardless of* X'
Conditional Probability	$P(Y = y \vert X = x)$	'The probability that $Y = y$ *given that* $X = x$'

Table : The different basic probability distributions.

A Pictorial Definition of Probability

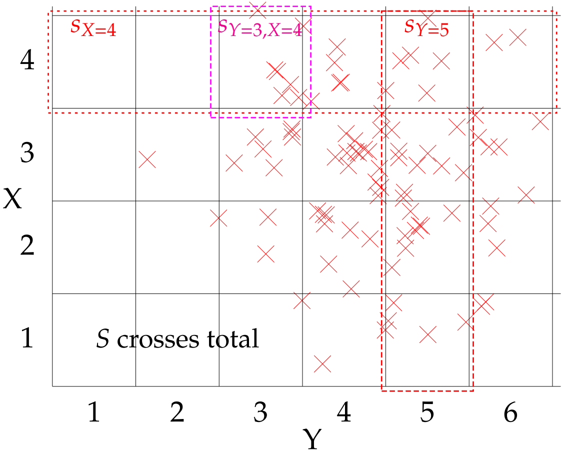


Figure : Representation of joint and conditional probabilities.

Different Distributions

Terminology	Definition	Notation
Joint Probability	$\lim_{S\to\infty} \frac{s_{Y=3,X=4}}{S}$	$P\left(Y=3, X=4\right)$
Marginal Probability	$\lim_{S\to\infty} \frac{s_{Y=5}}{S}$	$P\left(Y=5\right)$
Conditional Probability	$\lim_{S\to\infty} \frac{s_{Y=3,X=4}}{s_{X=4}}$	$P\left(Y=3 \vert X=4\right)$

Table : Definition of probability distributions.

Notational Details

- Typically we should write out $P(Y = y, X = x)$.
- In practice, we often use $P(y, x)$.
- This looks very much like we might write a multivariate function, *e.g.* $f(y, x) = \frac{y}{x}$.
  - For a multivariate function though, $f(y, x) \neq f(x, y)$.
  - However $P(y, x) = P(x, y)$ because $P(Y = y, X = x) = P(X = x, Y = y)$.
- We now quickly review the 'rules of probability'.

Normalization

All distributions are normalized. This is clear from the fact that $\sum_y s_y = S$, which gives

$$\sum_y P(y) = \frac{\sum_y s_y}{S} = \frac{S}{S} = 1.$$

A similar result can be derived for the marginal and conditional distributions.

The Sum Rule

Ignoring the limit in our definitions:

- The marginal probability $P(x)$ is $\frac{s_x}{S}$ (ignoring the limit).
- The joint distribution $P(y, x)$ is $\frac{s_{y,x}}{S}$.
- $s_x = \sum_y s_{y,x}$ so

$$\frac{s_x}{S} = \sum_y \frac{s_{y,x}}{S},$$

  in other words

$$P(x) = \sum_y P(y, x).$$

  This is known as the sum rule of probability.

The Product Rule

- $P(y|x)$ is
$$\frac{s_{y,x}}{s_x}.$$
- $P(y,x)$ is
$$\frac{s_{y,x}}{S} = \frac{s_{y,x}}{s_x}\frac{s_x}{S}$$
or in other words
$$P(y,x) = P(y|x)\,P(x).$$
This is known as the product rule of probability.

Bayes' Rule

- From the product rule,

$$P(x, y) = P(y, x) = P(y|x)\, P(x)\,,$$

so

$$P(x|y)\, P(y) = P(y|x)\, P(x)$$

which leads to Bayes' rule,

$$P(x|y) = \frac{P(y|x)\, P(x)}{P(y)}.$$

Outline

Basic Probability

Basic Probability

Probability Density Functions

Continuous Variables

- For continuous models we use the *probability density function* (PDF).
- Discrete case: defined probability distributions over a discrete number of states.
- How do we represent continuous as probability?
- Student heights:
  - Develop a representation which could answer *any* question we chose to ask about a student's height.
- PDF is a positive function, integral over the region of interest is one[1].

Manipulating PDFs

- Same rules for PDFs as distributions *e.g.*

$$p\left(x|y\right) = \frac{p\left(y|x\right)p\left(x\right)}{p\left(y\right)}$$

  where $p\left(y,x\right) = p\left(y|x\right)p\left(x\right)$ and for continuous variables $p\left(y\right) = \int p\left(y,x\right)\mathrm{d}x$.
- Expectations under a PDF

$$\left\langle f\left(y\right)\right\rangle_{p(y)} = \int f\left(y\right)p\left(y\right)\mathrm{d}y$$

  where the integral is over the region for which our PDF for y is defined.

The Gaussian Density

- Perhaps the most common probability density.

$$
\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\\
&= \mathcal{N}\left(x|\mu,\sigma^2\right)
\end{aligned}
$$

- Also available in multivariate form.
- First proposed maybe by de Moivre but also used by Laplace.

Gaussian PDF I

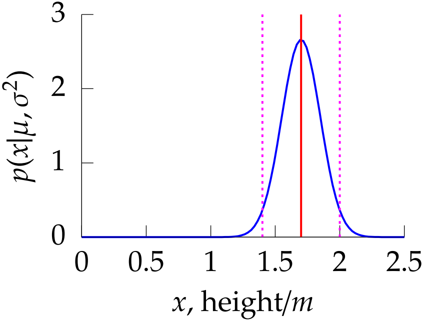


Figure : The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. Mean shown as red line. Two standard deviations are shown as magenta. It could represent the heights of a population of students.

Regression Revisited

- We introduced an error function of the form

$$E(\mathbf{w}) = \sum_{i=1}^{n} (y_i - mx_i - c)^2$$

- Quadratic error functions can be seen as Gaussian noise models.
- Imagine we are seeing data given by,

$$y(x_i) = mx_i + c + \epsilon$$

where ϵ is Gaussian noise with standard deviation σ,

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

Noise Corrupted Mapping

- This implies that

$$y_i \sim \mathcal{N}\left(mx_i + c, \sigma^2\right)$$

- Which we also write

$$p(y_i|\mathbf{w}, \sigma) = \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) \propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) \propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|m, c, \sigma^2) \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2}\right)$$

Gaussian Log Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$\log p(\mathbf{y}|m, c, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - c)^2 + \text{const}$$

Gaussian Log Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$-\log p(\mathbf{y}|m, c, \sigma^2) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - mx_i - c)^2 + \text{const}$$

Gaussian Log Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|m, c, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|mx_i + c, \sigma^2\right)$$

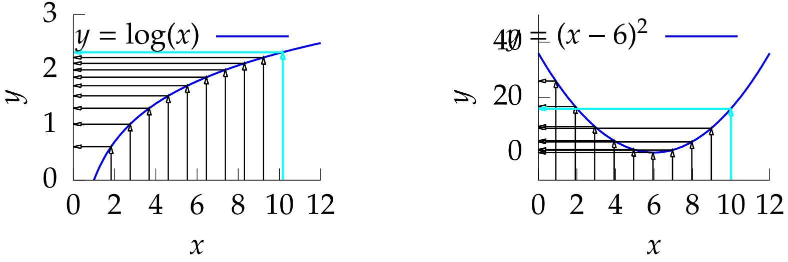
- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$-\log p(\mathbf{y}|m, c, \sigma^2) = \frac{1}{2\sigma^2} E(m, c) + \text{const}$$

Probabilistic Interpretation of the Error Function

- Probabilistic Interpretation for Error Function is Negative Log Likelihood.
- *Minimizing* error function is equivalent to *maximizing* log likelihood.
- Maximizing *log likelihood* is equivalent to maximizing the *likelihood* because log is monotonic.
- Probabilistic interpretation: Minimizing error function is equivalent to maximum likelihood with respect to parameters.

Monotonicity and Ordering



Monotonic functions preserve the ordering of input points, so the largest x is also the largest y. *Left*: gives an impression of this idea, cyan arrow is largest in x and correspondingly the largest in y. This transformation is log. *Right*: this quadratic function doesn't preserve the ordering and the largest x (again cyan arrow) is not the largest y value.

Sample Based Approximation implies i.i.d

- The log likelihood is

$$L(\boldsymbol{\theta}) = \log P(\mathbf{y}|\boldsymbol{\theta})$$

- If the likelihood is *independent* over the individual data points,

$$P(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{\theta})$$

- This is equivalent to the assumption that the data is *independent* and *identically* distributed. This is known as *i.i.d.*.
- Now the log likelihood is

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log P(y_i|\boldsymbol{\theta})$$

- We take the negative log likelihood to recover the sum of squares error.

References I

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, Cambridge, MA, 2002. MIT Press.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.