CS 2750 Machine Learning
Lecture 6

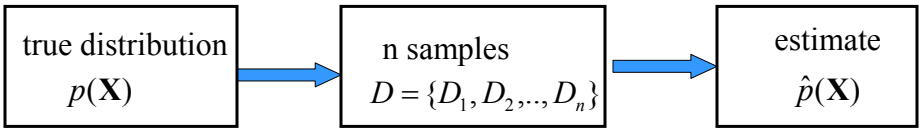
Density estimation III. Linear regression.

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$
$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: try to estimate the underlying 'true' probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D



Standard (iid) assumptions: Samples
- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**

Exponential family

Exponential family:

- all probability mass / density functions that can be written in the exponential normal form

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- $\boldsymbol{\eta}$ a vector of natural (or canonical) parameters
- $t(\mathbf{x})$ a function referred to as a sufficient statistic
- $h(\mathbf{x})$ a function of x (it is less important)
- $Z(\boldsymbol{\eta})$ a normalization constant (a partition function)

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}$$

- Other common form:

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})\right] \qquad \log Z(\boldsymbol{\eta}) = A(\boldsymbol{\eta})$$

Exponential family: examples

- **Bernoulli distribution**

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x}$$
$$= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}$$
$$= \exp\{\log(1-\pi)\}\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x\right\}$$

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- **Parameters**

$\boldsymbol{\eta} = ?$ $\qquad t(\mathbf{x}) = ?$

$Z(\boldsymbol{\eta}) = ?$ $\qquad h(\mathbf{x}) = ?$

Exponential family: examples

- **Bernoulli distribution**

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x}$$
$$= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}$$
$$= \exp\{\log(1-\pi)\}\exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x\right\}$$

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- **Parameters**

$$\boldsymbol{\eta} = \log\frac{\pi}{1-\pi} \quad (\text{note} \quad \pi = \frac{1}{1+e^{-\eta}}) \qquad t(\mathbf{x}) = x$$
$$Z(\boldsymbol{\eta}) = \frac{1}{1-\pi} = 1 + e^{\eta} \qquad h(\mathbf{x}) = 1$$

Exponential family: examples

- **Univariate Gaussian distribution**

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$
$$= \frac{1}{2\pi}\exp\left(-\frac{\mu}{2\sigma^2} - \log\sigma\right)\exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right\}$$

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(x) \exp\left[\eta^T t(x)\right]$$

- **Parameters**

$$\boldsymbol{\eta} = ? \qquad t(\mathbf{x}) = ?$$
$$Z(\boldsymbol{\eta}) = ? \qquad h(\mathbf{x}) = ?$$

Exponential family: examples

- **Univariate Gaussian distribution**

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu}{2\sigma^2} - \log\sigma\right) \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right\}$$

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(x) \exp\left[\eta^T t(x)\right]$$

- **Parameters**

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

$$Z(\boldsymbol{\eta}) = \exp\left\{\frac{\mu}{2\sigma^2} + \log\sigma\right\} = \exp\left\{-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right\}$$

$$h(\mathbf{x}) = 1/\sqrt{2\pi}$$

Exponential family

- **For iid samples, the likelihood of data is**

$$P(D \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} h(\mathbf{x}_i) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x}_i) - A(\boldsymbol{\eta})\right]$$

$$= \left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\sum_{i=1}^{n} \boldsymbol{\eta}^T t(\mathbf{x}_i) - A(\boldsymbol{\eta})\right]$$

$$= \left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\boldsymbol{\eta}^T \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

- **Important:**
  - the dimensionality of the sufficient statistic remains the same with the number of samples

Exponential family

- **The log likelihood of data is**

$$l(D,\boldsymbol{\eta}) = \log\left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\boldsymbol{\eta}^T\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

$$= \log\left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] + \left[\boldsymbol{\eta}^T\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

- **Optimizing the loglikelihood**

$$\nabla_{\boldsymbol{\eta}} l(D,\boldsymbol{\eta}) = \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - n\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) \quad = \mathbf{0}$$

- **For the ML estimate it must hold**

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \frac{1}{n}\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right)$$

Exponential family

- **Rewritting the gradient:**

Exponential family

- **Rewritting the gradient:**

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \log Z(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \log \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \frac{\int t(\mathbf{x}) h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}}{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \int t(\mathbf{x}) h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})\right\} d\mathbf{x}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = E(t(\mathbf{x}))$$

- **Result:** $$E(t(\mathbf{x})) = \frac{1}{n}\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right)$$
- **For the ML estimate the parameters $\boldsymbol{\eta}$ should be adjusted such that the expectation of the statistic t(x) is equal to the observed sample statistics**

Moments of the distribution

- **For the exponential family**
  - The k-th moment of the statistic corresponds to the k-th derivative of $A(\boldsymbol{\eta})$
  - If x is a component of t(x) then we get the moments of the distribution by differentiating its corresponding natural parameter
- **Example: Bernoulli** $p(x \mid \pi) = \exp\left\{\log\left(\frac{\pi}{1-\pi}\right)x + \log(1-\pi)\right\}$

$$A(\boldsymbol{\eta}) = \log\frac{1}{1-\pi} = \log(1+e^{\eta})$$

- **Derivatives:**

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta} = \frac{\partial}{\partial \eta}\log(1+e^{\eta}) = \frac{e^{\eta}}{(1+e^{\eta})} = \frac{1}{(1+e^{-\eta})} = \pi$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta^2} = \frac{\partial}{\partial \eta}\frac{1}{(1+e^{-\eta})} = \pi(1-\pi)$$

Outline

Linear Regression

- Linear model
- Error function based on the least squares fit
- Parameter estimation.

Supervised learning

Data: $D = \{D_1, D_2, .., D_n\}$ **a set of *n* examples**

$D_i =< \mathbf{x}_i, y_i >$

$\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots x_{i,d})$ is an input vector of size d

y_i is the desired output (given by a teacher)

Objective: learn the mapping $f : X \to Y$

s.t. $y_i \approx f(\mathbf{x}_i)$ for all $i = 1, .., n$

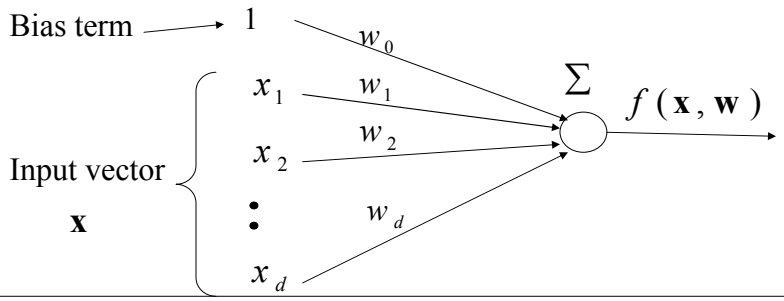
- **Regression:** Y is **continuous**
  Example: earnings, product orders → company stock price
- **Classification:** Y is **discrete**
  Example: handwritten digit in binary form → digit label

Linear regression

- **Function** $f : X \rightarrow Y$ is a linear combination of input components

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots w_d x_d = w_0 + \sum_{j=1}^{d} w_j x_j$$

$w_0, w_1, \ldots w_k$ **- parameters (weights)**



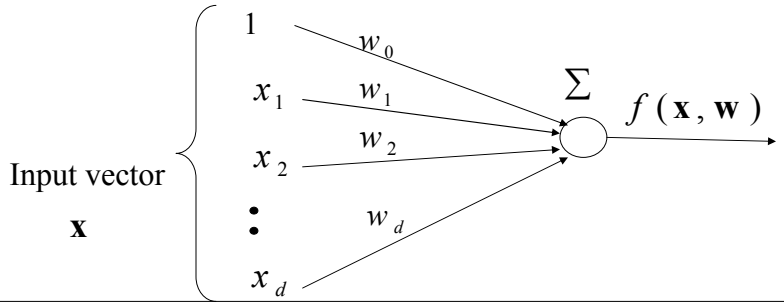
Linear regression

- **Shorter (vector) definition of the model**
  - Include bias constant in the input vector

$$\mathbf{x} = (1, x_1, x_2, \cdots x_d)$$

$$f(\mathbf{x}) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots w_d x_d = \mathbf{w}^T \mathbf{x}$$

$w_0, w_1, \ldots w_k$ **- parameters (weights)**



Linear regression. Error.

- **Data:** $D_i =< \mathbf{x}_i, y_i >$
- **Function:** $\mathbf{x}_i \rightarrow f(\mathbf{x}_i)$
- We would like to have $y_i \approx f(\mathbf{x}_i)$ for all $i = 1,.., n$

- **Error function**
  - measures how much our predictions deviate from the desired answers

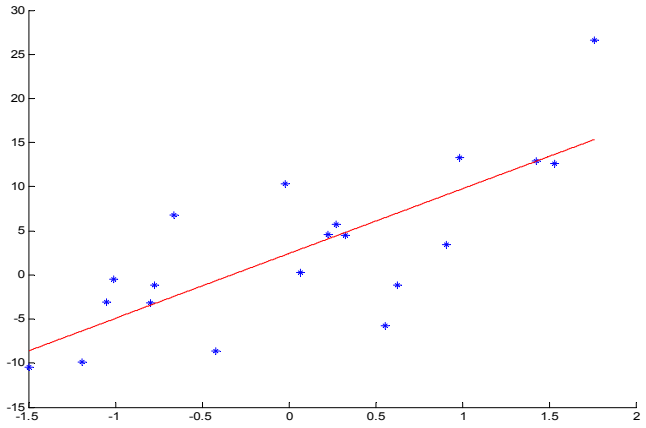
Mean-squared error $J_n = \frac{1}{n} \sum_{i=1,..n} (y_i - f(\mathbf{x}_i))^2$

- **Learning:**

  We want to find the weights minimizing the error !

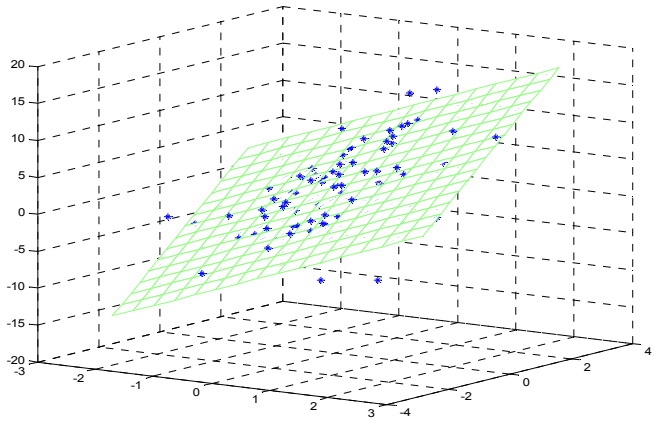
Linear regression. Example

- 1 dimensional input $\mathbf{x} = (x_1)$



Linear regression. Example.

- 2 dimensional input $\mathbf{x} = (x_1, x_2)$



Linear regression. Optimization.

- We want the **weights minimizing the error**

$$J_n = \frac{1}{n}\sum_{i=1,..n}(y_i - f(\mathbf{x}_i))^2 = \frac{1}{n}\sum_{i=1,..n}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

- For the optimal set of parameters, derivatives of the error with respect to each parameter must be 0

$$\frac{\partial}{\partial w_j}J_n(\mathbf{w}) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d})x_{i,j} = 0$$

- **Vector of derivatives:**

$$\text{grad}_{\mathbf{w}}(J_n(\mathbf{w})) = \nabla_{\mathbf{w}}(J_n(\mathbf{w})) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i = \overline{\mathbf{0}}$$

Linear regression. Optimization.

- $\text{grad}_{\mathbf{w}}(J_n(\mathbf{w})) = \overline{\mathbf{0}}$ defines a set of equations in $\mathbf{w}$

$$\frac{\partial}{\partial w_0} J_n(\mathbf{w}) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) = 0$$

$$\frac{\partial}{\partial w_1} J_n(\mathbf{w}) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,1} = 0$$

$$\ldots$$

$$\frac{\partial}{\partial w_j} J_n(\mathbf{w}) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,j} = 0$$

$$\ldots$$

$$\frac{\partial}{\partial w_d} J_n(\mathbf{w}) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,d} = 0$$

Solving linear regression

$$\frac{\partial}{\partial w_j} J_n(\mathbf{w}) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - w_0 x_{i,0} - w_1 x_{i,1} - \ldots - w_d x_{i,d}) x_{i,j} = 0$$

By rearranging the terms we get a **system of linear equations** with $d+1$ unknowns

$$\mathbf{Aw} = \mathbf{b}$$

$$w_0\sum_{i=1}^{n} x_{i,0}1 + w_1\sum_{i=1}^{n} x_{i,1}1 + \ldots + w_j\sum_{i=1}^{n} x_{i,j}1 + \ldots + w_d\sum_{i=1}^{n} x_{i,d}1 = \sum_{i=1}^{n} y_i 1$$

$$w_0\sum_{i=1}^{n} x_{i,0}x_{i,1} + w_1\sum_{i=1}^{n} x_{i,1}x_{i,1} + \ldots + w_j\sum_{i=1}^{n} x_{i,j}x_{i,1} + \ldots + w_d\sum_{i=1}^{n} x_{i,d}x_{i,1} = \sum_{i=1}^{n} y_i x_{i,1}$$

$$\bullet\bullet\bullet$$

$$w_0\sum_{i=1}^{n} x_{i,0}x_{i,j} + w_1\sum_{i=1}^{n} x_{i,1}x_{i,j} + \ldots + w_j\sum_{i=1}^{n} x_{i,j}x_{i,j} + \ldots + w_d\sum_{i=1}^{n} x_{i,d}x_{i,j} = \sum_{i=1}^{n} y_i x_{i,j}$$

$$\bullet\bullet\bullet$$

Solving linear regression

- The optimal set of weights satisfies:

$$\nabla_{\mathbf{w}}(J_n(\mathbf{w})) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i = \overline{\mathbf{0}}$$

Leads to a **system of linear equations (SLE)** with $d+1$ unknowns of the form

$$\mathbf{Aw} = \mathbf{b}$$

$$w_0\sum_{i=1}^{n}x_{i,0}x_{i,j} + w_1\sum_{i=1}^{n}x_{i,1}x_{i,j} + \ldots + w_j\sum_{i=1}^{n}x_{i,j}x_{i,j} + \ldots + w_d\sum_{i=1}^{n}x_{i,d}x_{i,j} = \sum_{i=1}^{n}y_i x_{i,j}$$

Solution to SLE: ?

Solving linear regression

- The optimal set of weights satisfies:

$$\nabla_{\mathbf{w}}(J_n(\mathbf{w})) = -\frac{2}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i = \overline{\mathbf{0}}$$

Leads to a **system of linear equations (SLE)** with $d+1$ unknowns of the form

$$\mathbf{Aw} = \mathbf{b}$$

$$w_0\sum_{i=1}^{n}x_{i,0}x_{i,j} + w_1\sum_{i=1}^{n}x_{i,1}x_{i,j} + \ldots + w_j\sum_{i=1}^{n}x_{i,j}x_{i,j} + \ldots + w_d\sum_{i=1}^{n}x_{i,d}x_{i,j} = \sum_{i=1}^{n}y_i x_{i,j}$$

Solution to SLE:

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{b}$$

- matrix inversion