CS 1571 Introduction to AI
Lecture 26

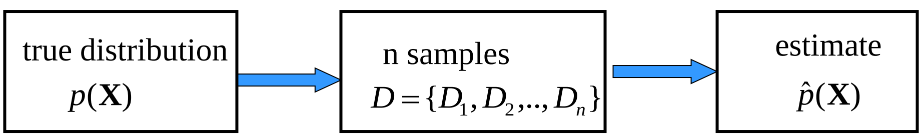
Learning probability distributions

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 $p(\mathbf{X})$)**

Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in ***X*** with parameters Θ
- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: find parameters $\hat{\Theta}$ that fit the data the best

- What is the best set of parameters?
  - There are various criteria one can apply here.

Parameter estimation. Basic criteria.

- **Maximum likelihood (ML)**

  maximize $p(D \mid \Theta, \xi)$

  ξ - represents prior (background) knowledge

- **Maximum a posteriori probability (MAP)**

  maximize $p(\Theta \mid D, \xi)$

  Selects the mode of the posterior

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)}$$

Maximum likelihood (ML) estimate.

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: D a sequence of outcomes such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

ML Solution: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

Maximum likelihood estimate. Example

- **Assume** the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of a head and a tail?

Maximum likelihood estimate. Example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10

What is the ML estimate of the probability of head and tail ?

Head: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$$

Tail: $$(1-\theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$

Maximum a posteriori estimate

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$$

Likelihood of data **prior**

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

$p(\theta \mid \xi)$ - is the prior probability on θ

How to choose the prior probability?

Prior distribution

Choice of prior: **Beta distribution**

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

$\Gamma(x)$ - A Gamma function

For integer values of x $\Gamma(x) = (x-1)!$

Why to use Beta distribution?

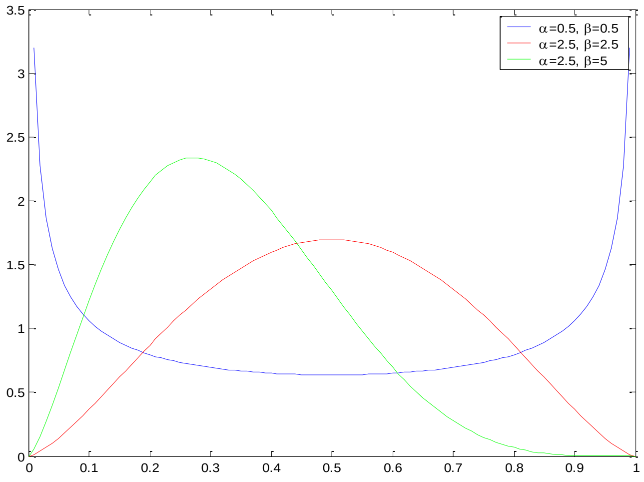
Beta distribution "**fits**" Bernoulli trials - **conjugate choices**

$$P(D \mid \theta, \xi) = \theta^{N_1} (1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

Beta distribution



Maximum a posterior probability

Maximum a posteriori estimate

– Selects the mode of the **posterior distribution**

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

MAP Solution: $$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

Note: that parameters of the prior

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_2 - 1}$$

– Act like counts of heads and tails
(sometimes they are also referred to as prior counts)

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate ?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- **The MAP can be highly biased with large prior counts**
- **It is hard to overturn it with a smaller sample size**
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T
  - **Heads:** 15
  - **Tails:** 10
- Assume

$$p(\theta \mid \xi) = Beta(\theta \mid 5,5) \qquad \theta_{MAP} = \frac{19}{33}$$

$$p(\theta \mid \xi) = Beta(\theta \mid 5,20) \qquad \theta_{MAP} = \frac{19}{48}$$

Multinomial distribution

Example: **Multi-way coin toss, roll of dice**

- **Data:** a set of N outcomes (multi-set)

  N_i - a number of times an outcome i has been seen

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$ **s.t.** $\sum_{i=1}^{k} \theta_i = 1$

θ_i - probability of an outcome i

Probability of data (likelihood)

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

MAP estimate

Choice of prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1, .., \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for multinomial

$$P(D \mid \boldsymbol{\theta}, \xi) = P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Posterior distribution

$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, .. \alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, .., \alpha_k + N_k)$$

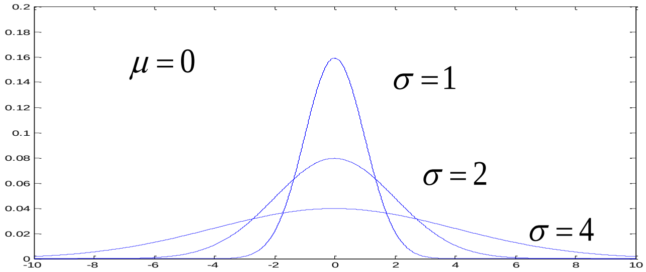
MAP estimate:

$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean
  σ - standard deviation
- **Density function:**

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



Parameter estimates

- **Log-likelihood** $l(D, \mu, \Sigma) = \log \prod_{i=1}^{n} p(x_i \mid \mu, \Sigma)$
- **ML estimates of the mean and covariances:**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

  – Covariance estimate is biased

$$E_n(\sigma^2) = E_n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- **Unbiased estimate:**

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Learning complex distributions

- **The problem of learning complex distributions**
  - can be sometimes reduced to the problem of learning a set of simpler distributions

- Such a decomposition occurs for example in **Bayesian networks**
  - Builds upon independences encoded in the network

- **Why learning of BBNs?**
  - Large databases are available
    - uncover important probabilistic dependencies from data and use them in inference tasks

Learning of BBN parameters

Learning. Two steps:
- Learning of the network structure
- Learning of parameters of conditional probabilities

- **Variables**:
  - Observable – values present in every data sample
  - Hidden – values are never in the sample
  - Missing values – values sometimes present, sometimes not

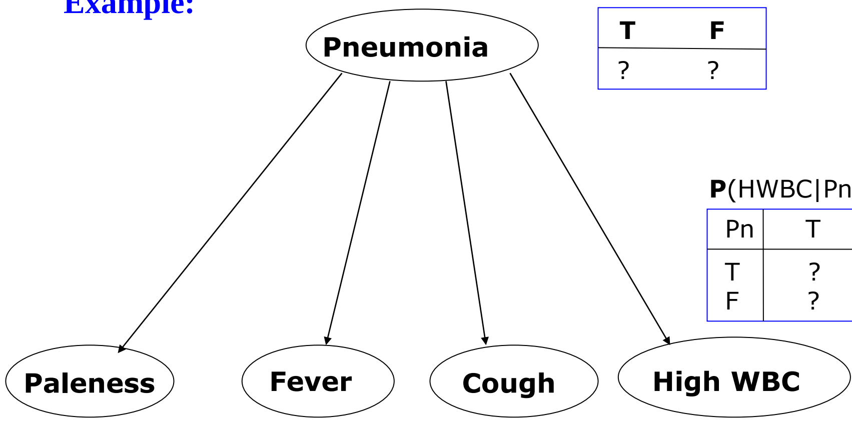
- **Here:**
  - learning parameters for the fixed graph structure
  - All variables are observed in the dataset

Learning of BBN parameters. Example.

Example:

P(Pneumonia)

T	F
?	?



P(HWBC|Pneum)

Pn	T	F
T	?	?
F	?	?

P(Palen|Pneum) ?

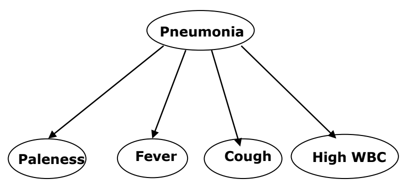
P(Fever|Pneum) ?

P(Cough|Pneum) ?

Learning of BBN parameters. Example.

Data D (different patient cases):

Pal	Fev	Cou	HWB	Pneu
T	T	T	T	F
T	F	F	F	F
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	F	T	F	F
F	F	F	F	F
T	T	F	F	F
T	T	T	T	T
F	T	F	T	T
T	F	F	T	F
F	T	F	F	F



Estimates of parameters of BBN

- Much like multiple **coin toss or roll of a dice** problems.
- A “smaller” learning problem corresponds to the learning of exactly one conditional distribution
- **Example:**

$$\mathbf{P}(Fever \mid Pneumonia = T)$$

- **Problem:** How to pick the data to learn?

Estimates of parameters of BBN

Much like multiple **coin toss or roll of a dice** problems.

- A “smaller” learning problem corresponds to the learning of exactly one conditional distribution

Example:

$$\mathbf{P}(Fever \mid Pneumonia = T)$$

Problem: How to pick the data to learn?

Answer:

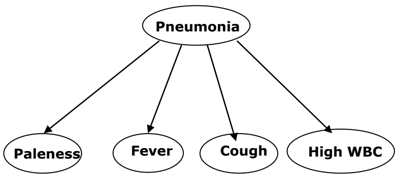
1. Select data points with Pneumonia=T (ignore the rest)
2. Focus on (select) only values of the random variable defining the distribution (Fever)
3. Learn the parameters of the conditional the same way as we learned the parameters of the biased coin or dice

Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 1: Select data points with Pneumonia=T

Pal	Fev	Cou	HWB	Pneu
T	T	T	T	F
T	F	F	F	F
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	F	T	F	F
F	F	F	F	F
T	T	F	F	F
T	T	T	T	T
F	T	F	T	T
T	F	F	T	F
F	T	F	F	F

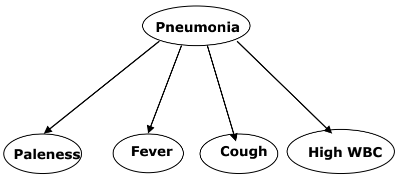


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 1: Ignore the rest

Pal	Fev	Cou	HWB	Pneu
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	T	T	T	T
F	T	F	T	T

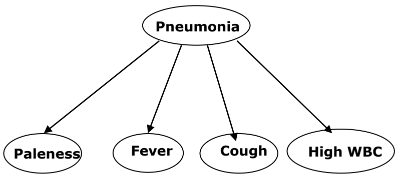


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 2: Select values of the random variable defining the distribution of Fever

Pal	Fev	Cou	HWB	Pneu
F	F	T	T	T
F	F	T	F	T
F	T	T	T	T
T	T	T	T	T
F	T	F	T	T

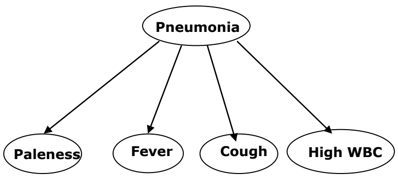


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 2: Ignore the rest

Fev
F
F
T
T
T

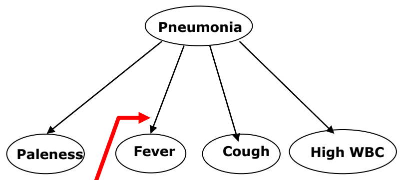


Learning of BBN parameters. Example.

Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 3: Learning the ML estimate

Fev
F
F
T
T
T



$\mathbf{P}(Fever \mid Pneumonia = T)$

T	F
0.6	0.4

Learning of BBN parameters. Example.

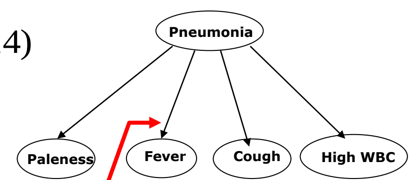
Learn: $\mathbf{P}(Fever \mid Pneumonia = T)$

Step 3: Learning the MAP estimate

Assume the prior

$$\theta_{Fever|Pneumonia=T} \sim Beta(3,4)$$

Fev
F
F
T
T
T



$\mathbf{P}(Fever \mid Pneumonia = T)$

T	F
0.5	0.5

Linear regression

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 $f : X \rightarrow Y$

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

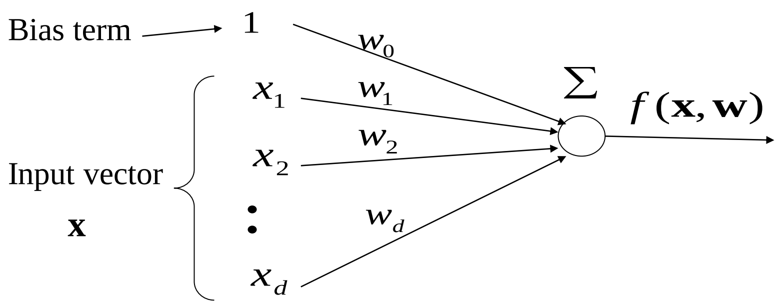
- **Regression:** Y is **continuous**
  Example: earnings, product orders $\rightarrow$ company stock price
- **Classification:** Y is **discrete**
  Example: handwritten digit in binary form $\rightarrow$ digit label

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)**



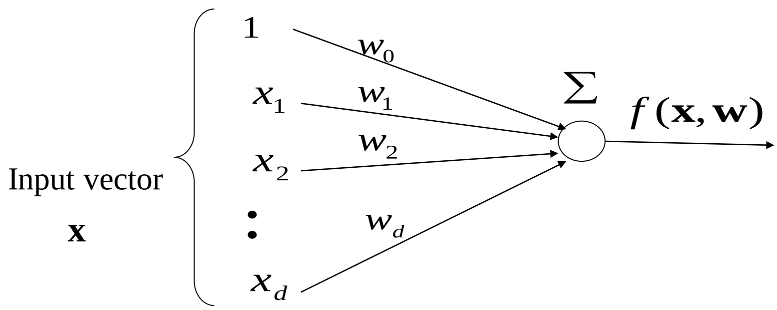
Linear regression

- **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 $i = 1,..,n$

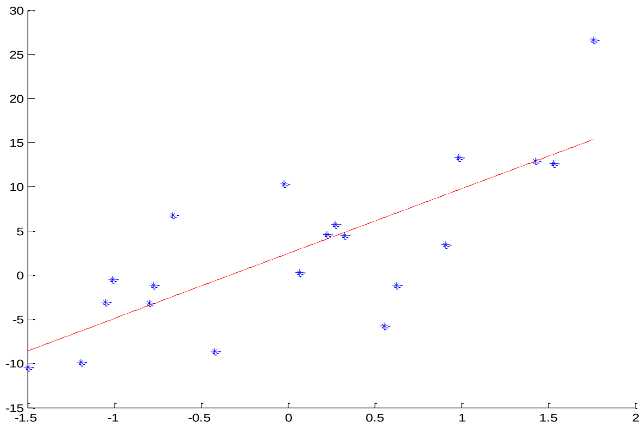
- **Error function** measures how much our predictions deviate from the desired answers

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 !

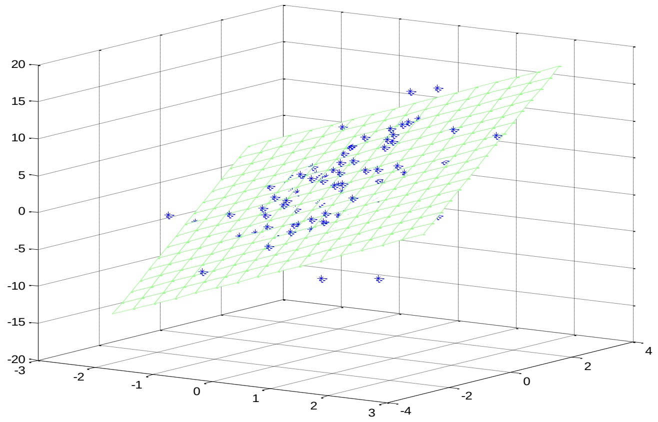
Linear regression. Example

- 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:**

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

Linear regression. Optimization.

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

$$J_n = \frac{1}{n}\sum_{i=1,..n}(y_i - f(\mathbf{x}_i))^2 = \frac{1}{n}\sum_{i=1,..n}(y_i - [w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots w_k x^{(k)}])^2$$

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

$$\frac{\partial}{\partial w_0} J_n(w) = -\frac{2}{n}\sum_{i=1}^{n}[y_i - (w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots w_k x^{(k)})] = 0$$

$$\frac{\partial}{\partial w_1} J_n(w) = -\frac{2}{n}\sum_{i=1}^{n}[y_i - (w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots w_k x^{(k)})]x^{(1)} = 0$$

$$\ldots$$

$$\frac{\partial}{\partial w_j} J_n(w) = -\frac{2}{n}\sum_{i=1}^{n}[y_i - (w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots w_k x^{(k)})]x^{(j)} = 0$$

$$\ldots$$

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

$$\mathbf{A}\mathbf{w} = \mathbf{b}$$

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

Solutions to SLE:

- e.g. matrix inversion (if the matrix is singular)

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

Gradient descent solution

- There are other ways to solve the weight optimization problem in the linear regression model

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

- A simple technique:
  - **Gradient descent**

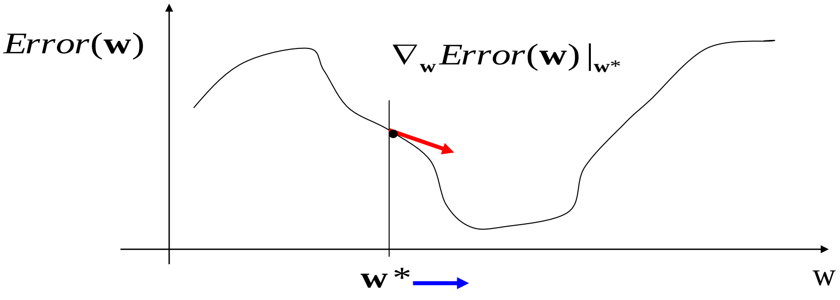
    Idea:
    - Adjust weights in the direction that improves the Error
    - The gradient tells us what is the right direction

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} Error_i(\mathbf{w})$$

$\alpha > 0$ - a learning rate (scales the gradient changes)

Gradient descent method

- Descend using the gradient information

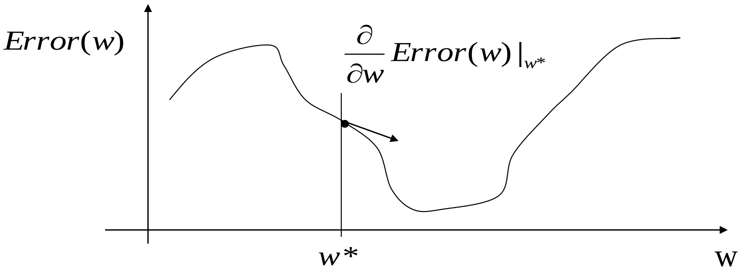


Direction of the descent

- Change the value of **w** according to the gradient

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} Error_i(\mathbf{w})$$

Gradient descent method



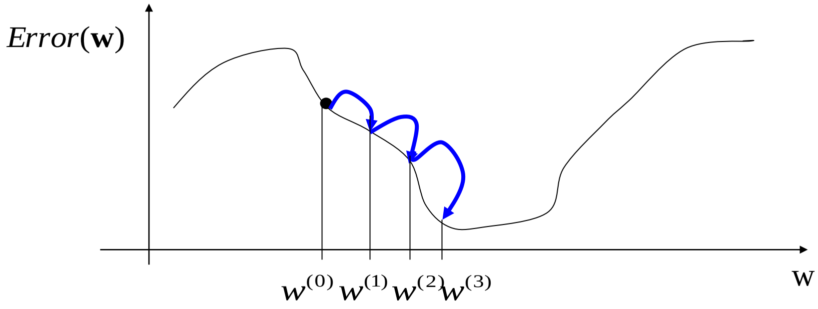
- New value of the parameter

$$w_j \leftarrow w_j^* - \alpha \frac{\partial}{\partial w_j} Error(w)|_{w^*} \qquad \text{For all j}$$

$\alpha > 0$ - a learning rate (scales the gradient changes)

Gradient descent method

- Iteratively converge to the optimum of the Error function



Online regression algorithm

- The error function defined for the whole dataset *D*

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

- Instead of the error for all data points we **use error for each exampls** $D_i = <\mathbf{x}_i, y_i>$

$$J_{\text{online}} = Error_i(\mathbf{w}) = \frac{1}{2}(y_i - f(\mathbf{x}_i, \mathbf{w}))^2$$

- **Change regression weights after every example according to the gradient:**

$$w_j \leftarrow w_j - \alpha \frac{\partial}{\partial w_j} Error_i(\mathbf{w})$$

vector form: $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} Error_i(\mathbf{w})$

$\alpha > 0$ - Learning rate that depends on the number of updates

Gradient for on-line learning

Linear model $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$

On-line error $J_{online} = Error_i(\mathbf{w}) = \frac{1}{2}(y_i - f(\mathbf{x}_i, \mathbf{w}))^2$

On-line algorithm: sequence of online updates

(i)-th update for the linear model: $D_i = < \mathbf{x}_i, y_i >$

Vector form:

$$\mathbf{w}^{(i)} \leftarrow \mathbf{w}^{(i-1)} - \alpha(i) \nabla_{\mathbf{w}} Error_i(\mathbf{w}) |_{\mathbf{w}^{(i-1)}} = \mathbf{w}^{(i-1)} + \alpha(i)(y_i - f(\mathbf{x}_i, \mathbf{w}^{(i-1)}))\mathbf{x}_i$$

j-th weight:

$$w_j^{(i)} \leftarrow w_j^{(i-1)} - \alpha(i) \frac{\partial Error_i(\mathbf{w})}{\partial w_j} |_{\mathbf{w}^{(i-1)}} = w_j^{(i-1)} + \alpha(i)(y_i - f(\mathbf{x}_i, \mathbf{w}^{(i-1)}))x_{i,j}$$

Annealed learning rate: $\alpha(i) \approx \frac{1}{i}$

- Gradually rescales changes in weights

Online regression algorithm

Online-linear-regression (*D, number of iterations*)

Initialize weights $\mathbf{w} = (w_0, w_1, w_2 \dots w_d)$

for i=1:1*: number of iterations*

do **select** a data point $D_i = (\mathbf{x}_i, y_i)$ from *D*

set $\alpha = 1/i$

update weight vector

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha(y_i - f(\mathbf{x}_i, \mathbf{w}))\mathbf{x}_i$$

end for

return weights $\mathbf{w}$

- **Advantages:** very easy to implement, continuous data streams

On-line learning.   Example