Selected topics in Data Science

Grandjean, Martin (2014).
"La connaissance est un réseau".
Les Cahiers du Numérique

Wojciech Krzemień

30.10 2020, NCBJ

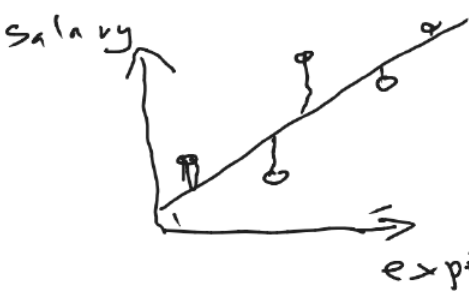
Last lecture recap

- Elements of Statistical Learning Theory:
  - loss function
  - empirical risk minimization
  - overfitting and model complexity
  - prediction vs inference

- Crash course in statistics I:
  - sample space, event space, probability measure
  - random variables, conditional probabilities and independent events
  - Bayes theorem
  - Cumulative Distribution Function (CDF), Probability Mass Function (PMF), Probability Density Function (PDF)

Loss function L to penalize the mistakes the model makes

regression



$\{x^1, y^1\}, \{x^2, y^2\} \ldots$

$L : X \times Y \to [0, \infty)$

quadratic loss → $L(y^i, f(x^i)) \longrightarrow (y^i - f(x^i))^2$

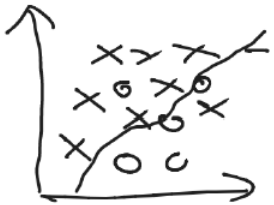
Mean squared error

training error → $E_T = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)) \longrightarrow MSE = \frac{1}{N} \sum (y^i - f(x^i))^2$

$= \frac{1}{N} \sum |y^i - f(x^i)|$

L1, L2 norms

classification



$E_T = \frac{1}{N} \sum L(f(x^i), y^i)$

Indicator function → $\begin{cases} 0 \\ 1 \end{cases} \Sigma[f(x^i) \neq y^i]$

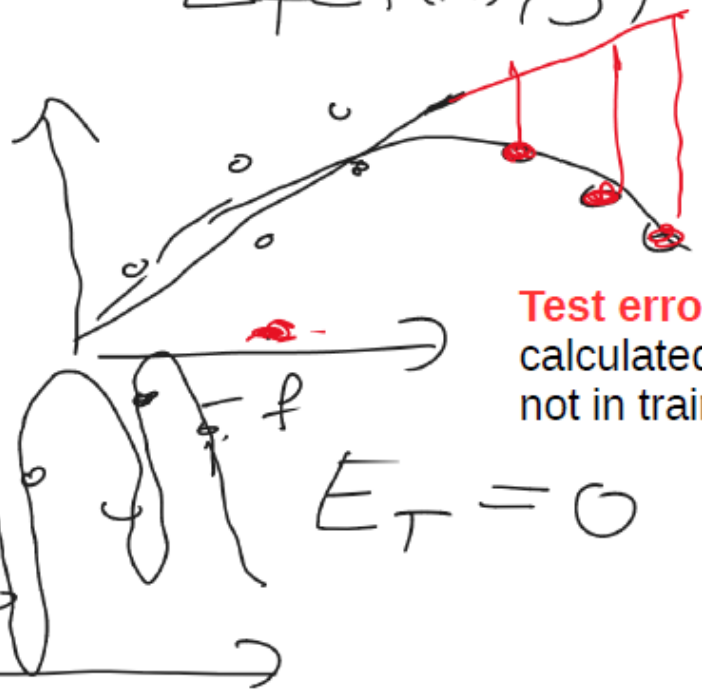
Training expressed as **minimization** problem

$f_\theta = \theta_0 + \theta_1 x_1$

Find parameters θ^* for which E_T has minimum

$E_T[f(x), y)$

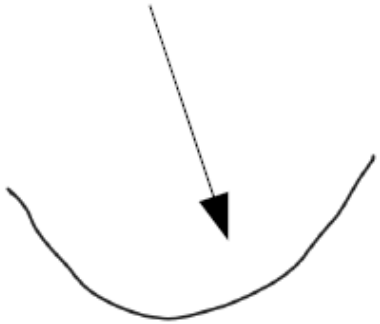
f_{θ^*}



Test errors
calculated for points
not in training set

OVERFITTING

Convex function



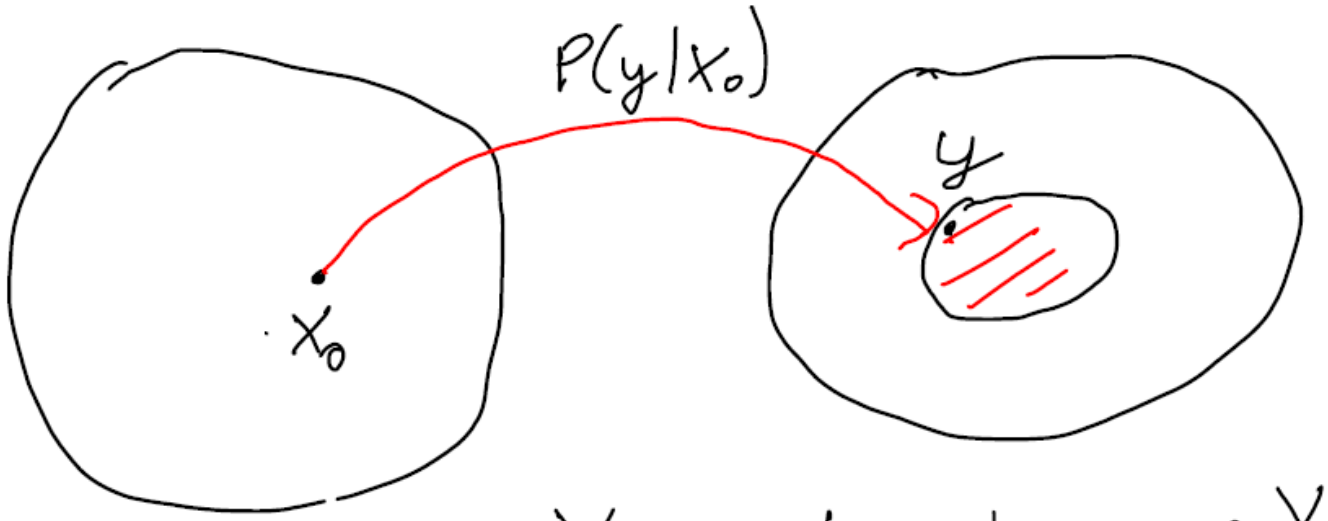
No more than one minimum

Generalization
Model should describe all the data

Statistical Learning Theory

$L : Y \times Y \to [0, \infty)$

Loss function



feature space X

e.g. $\mathbb{R}^p$

target space Y

e.g. $y \in \mathbb{R}$ - regression

$y \in \{0, 1\}$ - binary classification

$Z = X \times Y$, Z is a random variable with a joint probability density distribution

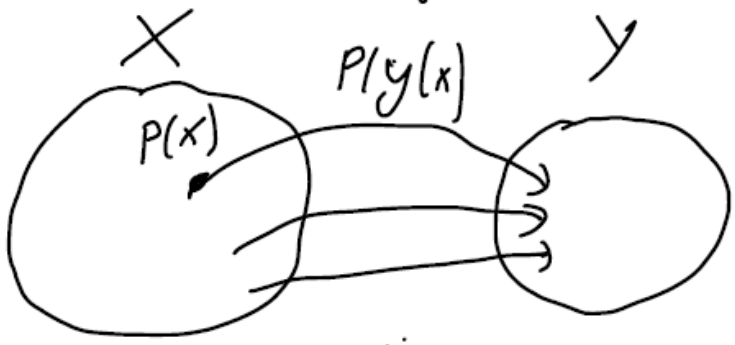
$$S_Z = S_{XY}$$

Expected prediction error (true error, risk function).

Joint probability density function

$$E_{PE}[f] = \int\int L(f(x), y)\, \rho_{X,Y}\, dx\, dy$$

The error that given model f would commit calculated over the whole $Z = X \times Y$ space.



Since we don't know what is $\rho_{X,Y}$ we cannot compute it

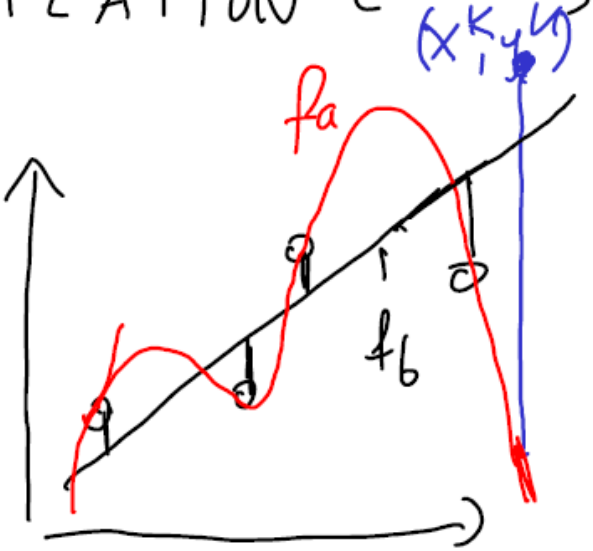
$$E_T[f] = \frac{1}{N}\sum_i^N L(f(x^i, y^i)) \quad \text{- training error, empirical risk}$$

For the 'good' model we expect: $E_T \approx E_{PE}$

GENERALIZATION ⟷ STABILITY

$E_T[f_a] = 0$

$E_T[f_b] > 0$



model complexity:

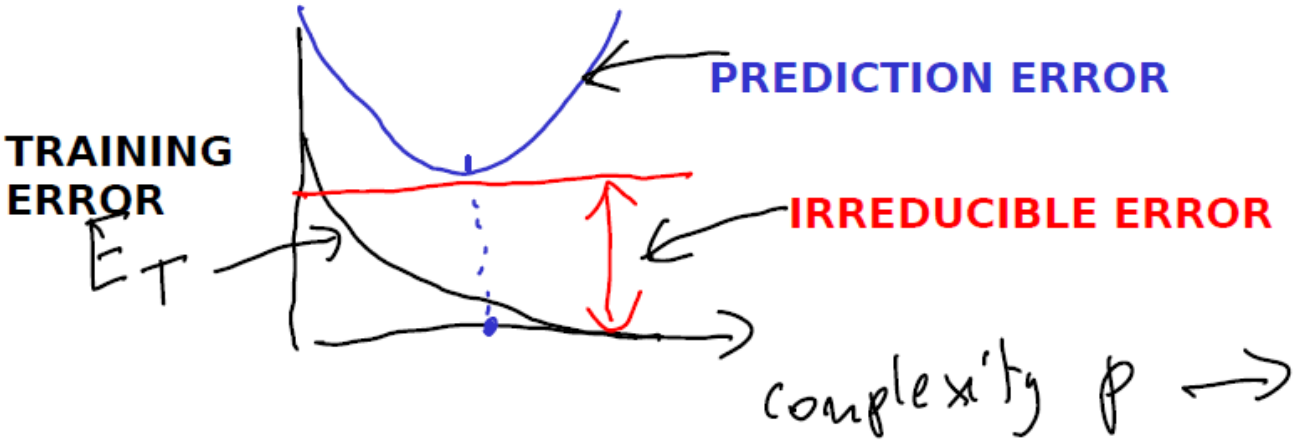
$f_1(x) = \theta_0 + \theta_1 x$

$f_2(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

⋮

$E[f_a(x^k, y^k)] \gg L[f_b(x^k, y^k)]$ violation of polynomial

↓ complexity



Todays plan

- Optimal algorithm and optimal classifier
- Bias-variance decomposition,
- k-NN revisited
- Crash course in Statistics II
- Work in groups:
  - Statistics problems
  - Small programs and k-NN implementations

Supervised vs unsupervised learning (Quiz)

Decide in each situation which kind of method would be better suited

1) A database with prices of the houses and its descriptions (e.g. area, location number of rooms). We want to predict prices for newly constructed houses.

2) A client database of the online shop. We want to group the customers according to their preferences.

3) A movie database with client profiles and a list of movies they watched, we want to develop a model to recommend movies for a given type of clients.

4) Samples of Internet traffic (e.g. packets), we would like to detect anomalous behaviours.

Supervised vs unsupervised learning (Quiz)

Decide in each situation which kind of method would be better suited

1) A database with prices of the houses and its descriptions (e.g. area, location number of rooms). We want to predict prices for newly constructed houses.

Supervised Learning

2) A client database of the online shop. We want to group the customers according to their preferences.

Unsupervised Learning

3) A movie database with client profiles and a list of movies they watched, we want to develop a model to recommend movies for a given type of clients.

Supervised Learning

4) Samples of Internet traffic (e.g. packets), we would like to detect anomalous behaviours.

Unsupervised Learning

Inference vs Prediction (Quiz)

Decide in each situation whether we are more interested in inference or in prediction.

1) The company launched an advertising campaign using different media e.g. TV, radio, or newspapers. The gathered data contains the sums of money invested in each sector and the rise in the income for different combinations. The boss would like to know what is the most optimal strategy e.g. is it worth advertising in all the sectors, is there any relation between the advertising on TV and radio, etc.

2) The company gathers data of their clients (e.g. age, education, preferences, and a lot more demographic information) and they want to promote their new product by a targeted mail campaign. The question is to decide which group of clients should be chosen.

Inference vs Prediction (Quiz)

Decide in each situation whether we are more interested in inference or in prediction.

1) The company launched an advertising campaign using different media e.g. TV, radio, or newspapers. The gathered data contains the sums of money invested in each sector and the rise in the income for different combinations. The boss would like to know what is the most optimal strategy e.g. is it worth advertising in all the sectors, is there any relation between the advertising on TV and radio, etc.

Inference / Both

2) The company gathers data of their clients (e.g. age, education, preferences, and a lot more demographic information) and they want to promote their new product by a targeted mail campaign. The question is to decide which group of clients should be chosen.

Prediction

Inference vs Prediction (Quiz)

Decide in each situation whether we are more interested in inference or in prediction.

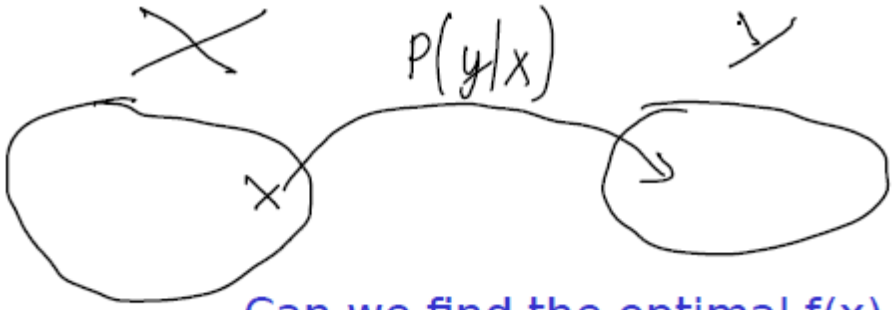
3) We analyze data from the university containing the results of students' exams together with the information about students and the courses that they previously took. We would like to understand what is the relation between the final scores they got and the other factors

Inference vs Prediction (Quiz)

Decide in each situation whether we are more interested in inference or in prediction.

3) We analyze data from the university containing the results of students' exams together with the information about students and the courses that they previously took. We would like to understand what is the relation between the final scores they got and the other factors

Inference



$$p(x,y) = p(y|x) \cdot p(x)$$

$$\varrho_{x,y} = \varrho_{y|x} \cdot \varrho_x$$

Can we find the optimal f(x), given quadratic loss function?

$$L[f(x), y] = (f(x) - y)^2$$

$$EPE[f] = \int\int L(f(x), y) \cdot \varrho_{x,y} \, dx\,dy$$

$$\int\int (f(x) - y)^2 \varrho_{y|x} \cdot \varrho_x \, dx\,dy =$$

$$\int_X \varrho_x \int_Y (f(x) - y)^2 \varrho_{y|x} \cdot dx\,dy$$

$$E[g(x)] \stackrel{df}{=} \int g(x) \varrho_x \, dx$$

$$\int p_x \int (f(x)-y)^2 p_{y|x}(y|x)\,dy\,dx$$

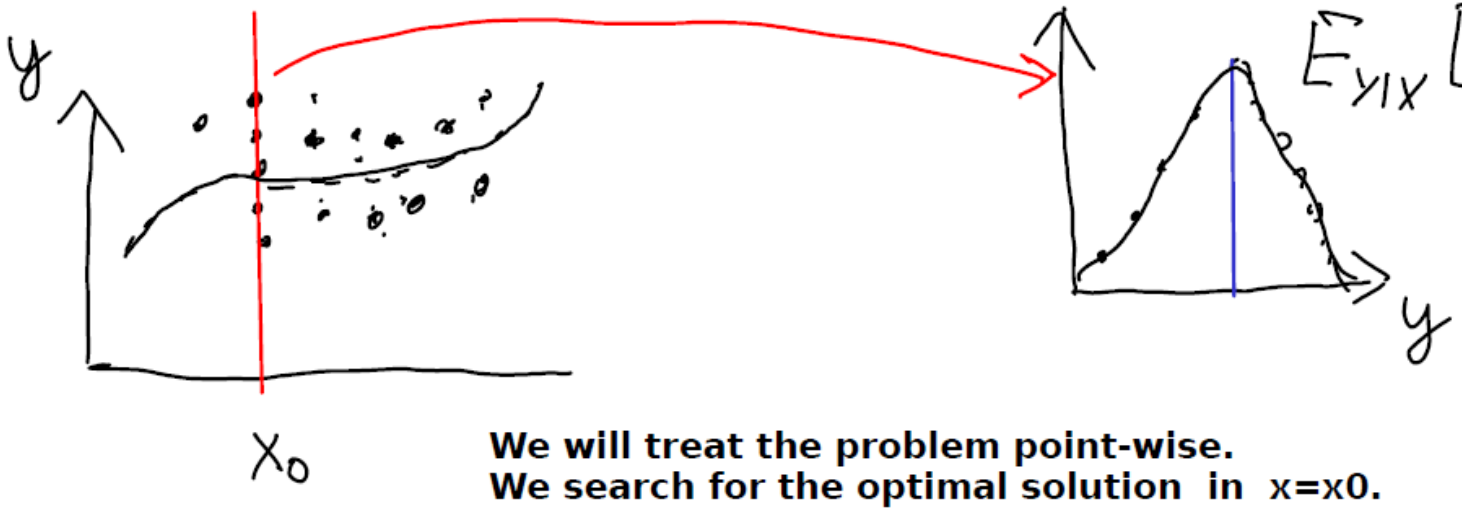
Expectation value over the whole feature space

$$E_x[E_{y|x}[(f(x)-y)^2|x]]$$

It can be approximated by the average for whole y's points.

$$E_{y|x}[y|x=x_0] \approx \frac{1}{N}\sum_{i=1}^{N} y_i$$

Example of the conditional expectation



We will treat the problem point-wise.
We search for the optimal solution in x=x0.

$$L = E_{y|x}[(f(x_0)-y)^2|x=x_0]$$

$$f(x_0) = c \quad \text{constant}$$

We want to find the minimum of L with respect to c

$$\frac{d}{dc}L = 0 \qquad \frac{d}{dc}E_{y|x}[(c-y)^2|x=x_0] = 0$$

$$C = f^* = E_{y|x}[y \mid x = x_0]$$ Regression Function

This is an optimal function for regression assuming MSE loss (!)

If we take $L(f(x), y) = |f(x) - y|$ one can find that minimization of EPE[f] gives $f^* = \text{Median}(y, X = x_0)$

Classification

Indicator function loss

$$L(f(x), y) = [f(x) \neq y]$$

$$\begin{cases} 1 \text{ if } f(x) \neq y \\ 0 \text{ if } f(x) = y \end{cases}$$

Again we're trying to find 'the optimal classifier' that minimizes EPE for indicator loss

$$EPE[f] = \int\int L(f(x), y)\, p_{y,x}\, dy\, dx$$

This is the classification problem so:

Discrete output over k classes:

$\int_y \cdot dy \rightarrow \sum_{k=1}^{K}$ sum over all output classes

$p_{y|x} \rightarrow p(k|x)$ e.g. for binary $y \in \{0,1\}$ classifier $\sum_{i=1}^{2}$

$$EPE[f] = E_X\left[\sum_{k=1}^{K} p(k|x) \cdot [f(x) \neq y]\right]$$

$$[f(x) \neq y] = 1 - [f(x) == y]$$

(...) After treating the problem point-wise $X = x_0$, we get the solution

$$f^* = \arg\max_k P(k|X = x_0)$$

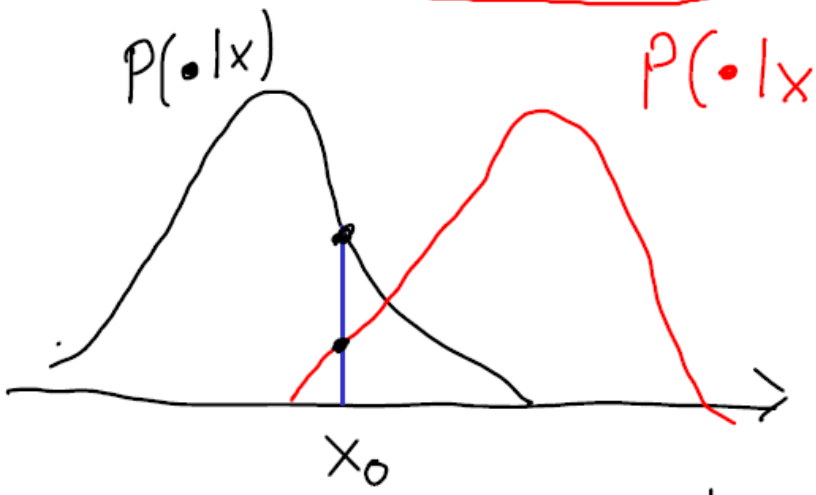
BAYES CLASSIFIER

Bayes classifier is "the optimal classifier" -> with the smallest error rate.

$$f_* = \arg\max_k P(k \mid X = x_0)$$

argmax [x] for f(x) is the notation to express x value for which f(x) has the maximum value:

e.g $\arg\max_x \sin(x) = \frac{\pi}{2}$



e.g $P(\bullet \mid x = x_0) = 70\%$
$P(\bullet \mid x = x_0) = 30\%$
→ We choose the class with higher probability so • . For 100 cases we commit 30% errors. (misclassifications)

$E_{rate} = 30\%$ but it's still the lowest possible !!!

Remark 1: Don't mix the Bayes classifier with the Naive Bayes Classifier.

Remark 2: Since normally we don't know P(y|x) we cannot use it.

Optimal f* model (algorithm)

- Let's assume that we know $\boldsymbol{\rho_{XY}(x,y)}$
- $f^*(X)$ – optimal algorithm for quadratic loss function $L(f(X), Y) = (f(X) - Y)^2$
- One can show that the optimal solution is:

$$\boldsymbol{f^*(X) = E_{Y|X}[Y|X]}$$

Problem:

Show that $f^*(X) = E_{Y|X}[Y|X]$ is optimal assuming $L(f(X), Y) = (f(X) - Y)^2$

$EPE(f) = E_{X,Y}[L(f(X),Y)] = \iint L(f(x),y)\, \rho_{XY}(x,y)dxdy$

Hints:

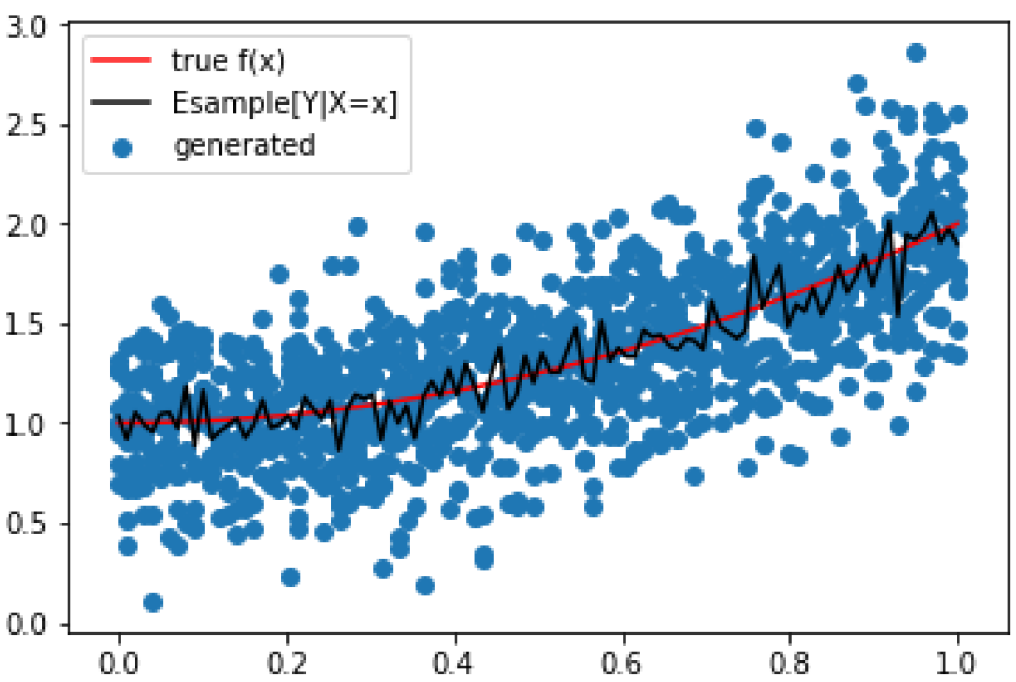
1.Express ρ_{XY} as conditional $\rho_{Y|X}$

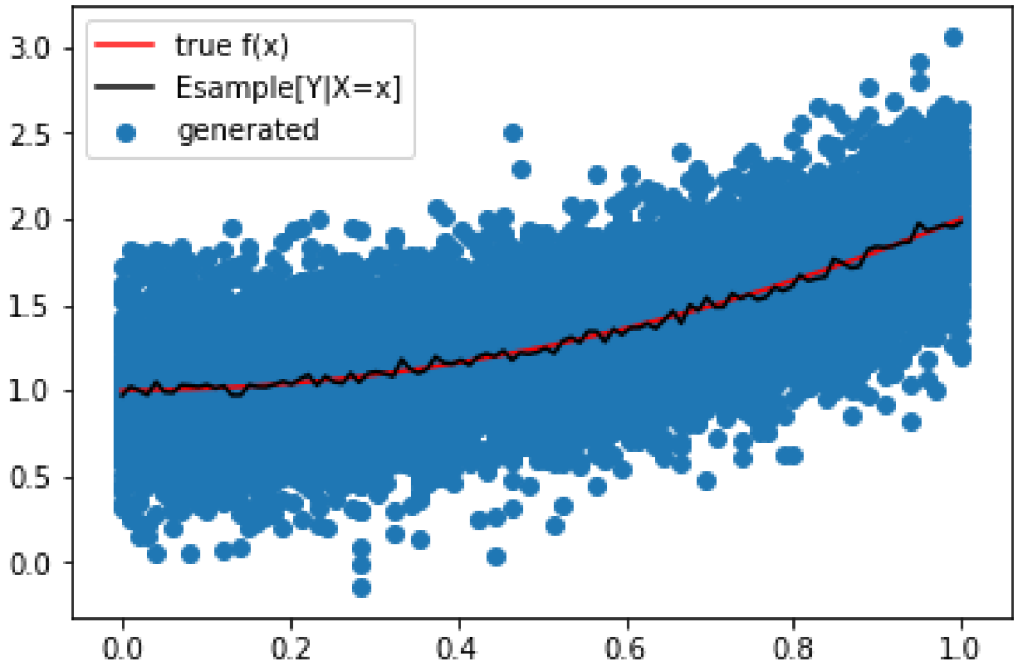
2. Express EPE as conditional expectations

3.Minimize the expression point-wise (for given $X = x_0$ from the feature space)

(*) regression function defined point-wise for every point x_0 of feature space X

Optimal f* model (regression function)





$$f^*(X) = E_{Y|X}[Y|X]$$

(*) regression function defined point-wise for every point x_0 of feature space X

Problem:

- Derive the optimal algorithm $f^*(X)$ assuming $L(f(X), Y) = |(f(X) - Y)|$

Problem:

- Derive the optimal algorithm $f^*(X)$ assuming $L(f(X), Y) = (f(X) - Y)^2$

Minimizing EPE[f]

L_2 $L = (f(x) - y)^2$

L_1: $L = |f(x) - y|$

Indicator loss for classification

$L = [f(x) \neq y]$

$f^* = E_{y|x}[y|x = x_0]$

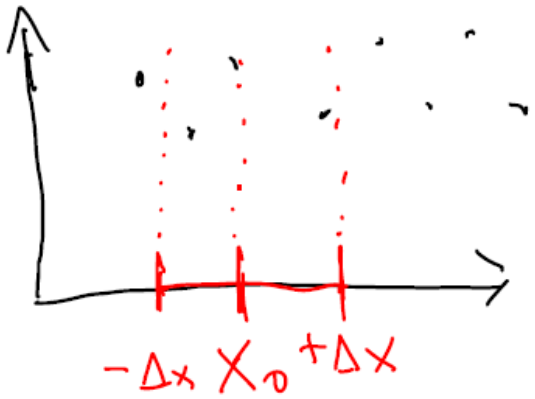
REGRESSION FUNCTION

$f^* = \text{Median}(y|x_0)$

$f^* = \underset{k}{\text{avg max}}\, P(k|x = x_0)$

BAYES CLASSIFIER

All f^* are defined point-wise, assuming that we know $g_{y,x}$ or $p(y,x)$.

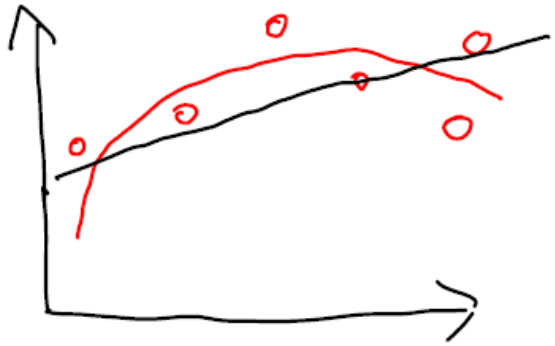


$f^* = E_{y|x}[y|x = x_0] \rightarrow$ Average y^i

$x^i \in [x - \Delta x, x + \Delta x]$

The k-NN for regression is the approximation of $E_{y|x}[y|x = x_0]$

Bias - Variance decomposition



$\hat{f}(x)$ - our model $= \theta_0 + x \cdot \theta_1$

$f^* = E_{y|x}[y|x]$ - optimal one

$y = f^*(x) + \varepsilon$ ← noise from Gaussian

$E[\varepsilon] = 0$

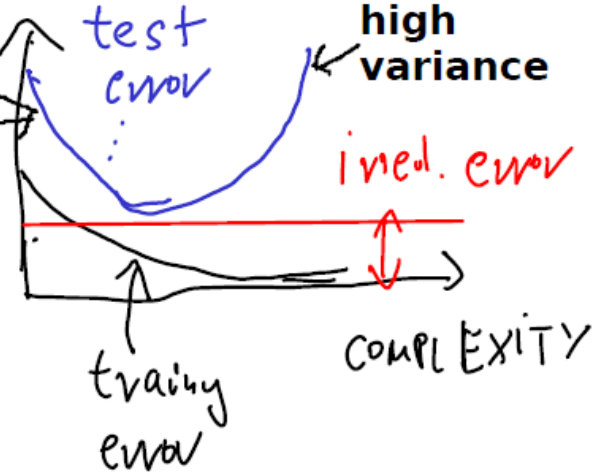
$Var[\varepsilon] = \sigma^2$

We decompose the error for given $X = x_0$

$$E_Y\left[\left(Y - \hat{f}(x_0)\right)^2 \mid X = x_0\right]$$

$$= \text{(show it!)}$$

$$VAR(\varepsilon) + BIAS^2(\hat{f}) + VAR(\hat{f})$$



Irreducible error

$(\hat{f} - f^*)^2$

How far we are from the optimal solution

How much or model would change for different training sets

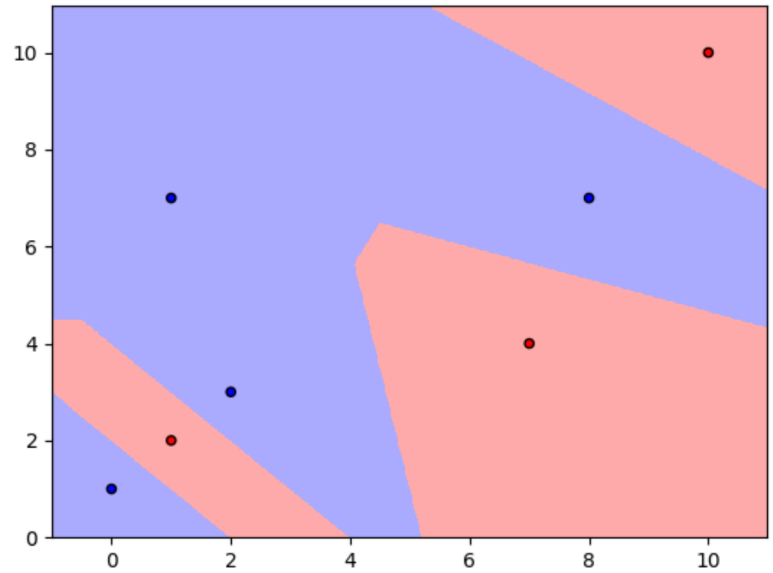
Notebook:

https://github.com/wkrzemien/dataScienceAndML2020/blob/master/notebooks/slt/EPE_train_test_errors.ipynb

k-Nearest Neighbors –model

X – space of our features, Y – space of outputs/labels of classes:

- Training set T consists of pairs of $\{X^{(i)}, Y^{(i)}\}$ i=1,...N
- Every X is feature vector with P coordinates $X=[x_1, x_2, \ldots, x_P]$
- k- number of Neighbors
- dist(X^i, X^j) is a distance metric in feature space X



To classify a new X^* instance:

- we choose k closest neighbors from T e.g. $N_k(x)$
- We calculate the average output Y (e.g. majority vote)
  $y(x) = 1/k * \sum Y^{(j)}$ over j for which $x^{(j)}$ belong to $N_k(x)$
- We assign (predict) the label of X*

- Problem Let's assume that we use kNN model to perform the classification of two type of objects (think of spam/non-spam) with k=2, k =3 and k=10 and N=100 training samples with 2 features each e.g. { $X^{(1)}=[0,1]^T$, $X^{(2)}=[-1,2]^T$, $X^{(3)}=[2,2]^T$, $X^{(4)}=[0,0.3]^T$, ..., $X^{(100)}=[-100,12,5]^T$ } and {$Y^{(1)}=[0]$, $Y^{(2)}=[1]$, $Y^{(3)}=[1]$, $Y^{(4)}=[0]$]

  - For which k value do we expect the smallest training error?
  - For which k value do we expect the highest/smallest stability?
  - How would we classify X=[0,0] if we use first 4 training samples for k=1, k=2, k=3? Calculate the majority vote result for each case.
  - We increase the training sample to N=101. How we expect to affect the stability for different k values?
  - Now instead of kNN model we use the linear regression. Discuss the stability issue.

Crash course in probability II

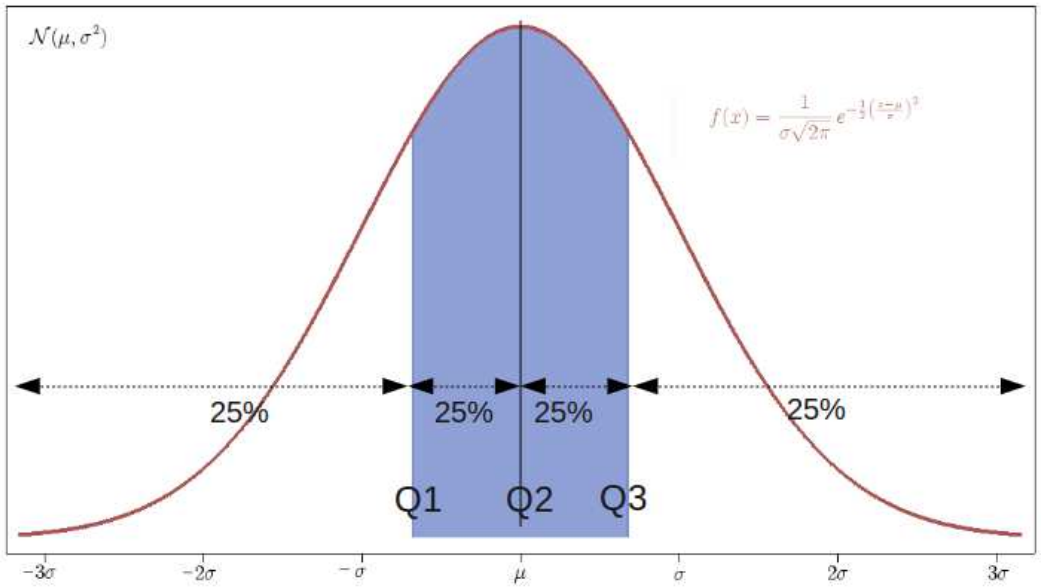
Quantiles

Quantile of the order q (q-quantile) x_q

cut points dividing the sample (distribution) into sub samples with equals probabilities:

$$F(x_q) = q \text{ for } q \text{ in } [0,1] \quad \rightarrow F^{-1}(q) = x_q \text{ if (cdf exists)}$$

Median - $x_{1/2}$ for which there is an equal probability that x gets lower and higher value.



Expectation / Expected value

Let g(x) be some arbitrary function: R → R, X – random variable, g(X) becomes random variable as well:

- For X being a discrete random variable with PMF $p_X(k)$:

$E[g(X)] = \Sigma\, p_X(k)\, g(k)$ for all possible k from X

- For X being a continuous random variable with PDF $f_X(x)$:

$E[g(X)] = \int g(x)\, f_X(x)dx$ for x $-\infty$ to ∞

Properties:

- **E[f(X) +g(X)] = E[f(X)] + E[g(X)]** - linearity
- E[K] = K for K being constant
- E[K* f(X)] = K *E[f(X)] for K being constant

- **Intuition:** E[g(X)] is a kind of weighted average with weights defined by p_x or f_x
- **Remark:** The equations for E[g(X)] are also known as the law of the unconscious statistician (LOTUS)

Variance

Variance of a random variable X expresses its concentration around the mean:

$$\mathbf{Var[X] = E[(E[X] - X)^2] \rightarrow Var[X] = E[X^2] - E^2[X]}$$

Properties:

- Var[K] = 0 for K being constant
- $Var[K * f(X)] = K^2 * Var[f(X)]$ for K being constant

Problem Let $g(X) = 1$ for some set A being a subset of sample space Ω:

- What is E[g(X)] if X is discrete with a given PMF or continuous with a given PDF

Problem

- What is E[g(X)] for g(X) =x

Problem

- Show that $Var[X] = E[X^2] - E^2[X]$

Problem

- Calculate the mean and the variance for the uniform distribution

Problem:

- Implement a function that returns mean
- Implement a function that returns Var[X]

Problem: Similarity measures

- Implement a function that returns Euclidean distance between two vectors
- Implement a function that returns Manhattan distance between two vectors

Joint and marginal CDF

Let's consider two random variables X, Y with CDF's: $F_X(x)$, and $F_Y(y)$, The joint CDF of X and Y is defined as:

$$\mathbf{F_{XY}(x,y) = P(X<=x, Y<=y)}$$

F_X and F_Y are called marginal CDFs and are related to F_{XY}

- $\mathbf{F_X(x) = \lim y \to \infty \ F_{XY}(x,y)}$
- $\mathbf{F_y(y) = \lim x \to \infty \ F_{XY}(x,y)}$

Remark: Since joint CDF is defined as probability, then obviously its values must be limited to range [0, 1].

Joint and marginal PMF and PDF

Let's consider two random variables X, Y with:

- (discrete) PMF's: $p_X(k)$, and $p_Y(r)$, The joint PMF is defined as:

  $$\mathbf{p_{XY}(k,r) = P(X=k, Y=r)}$$

- (continuous) with F_{XY} differentiable for all x,y values. The joint PDF is defined as:

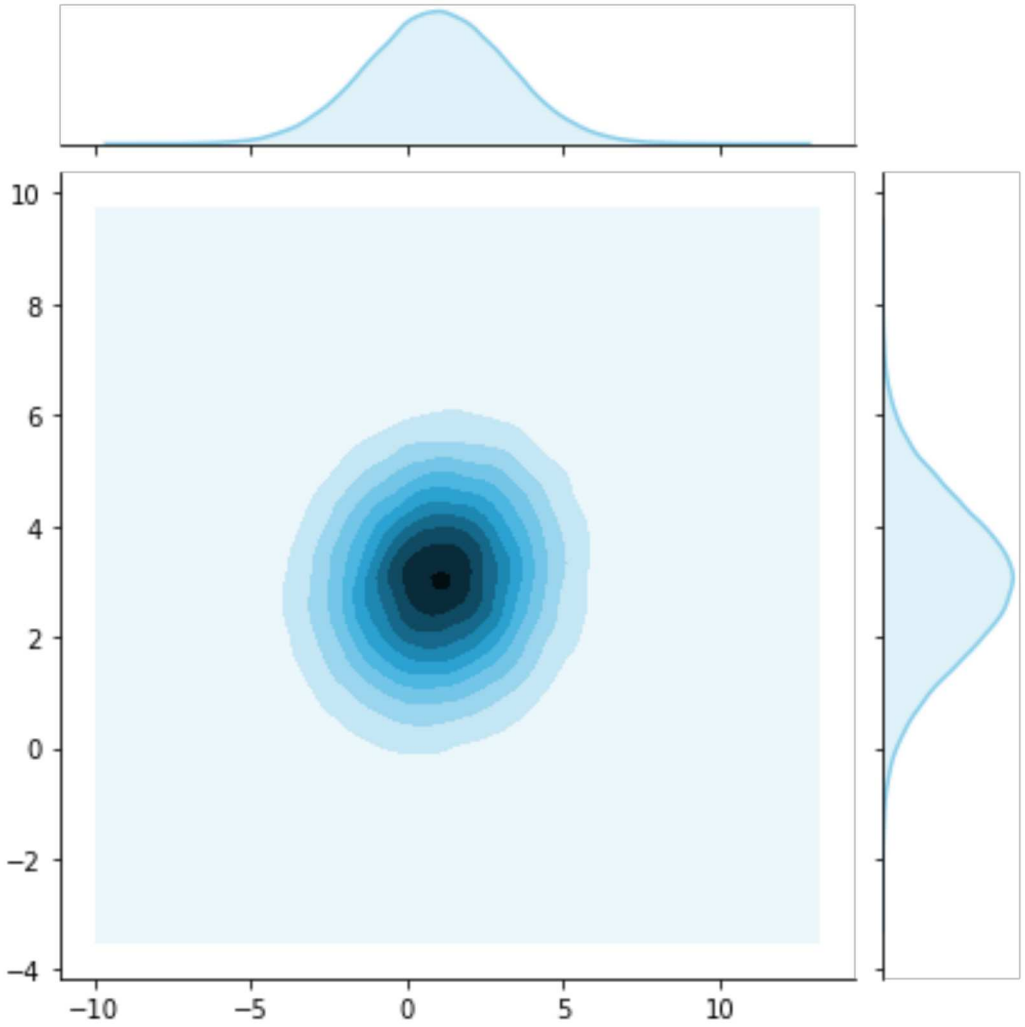
  $$\mathbf{f_{XY}(x,y) = \partial^2 F_{XY}(x,y)/\partial x \partial y}$$

  $$\mathbf{\iint f_{XY}(x,y)dxdy = P((X,Y) \in A) \text{ for } x,y \in A}$$

p_X /f_X and p_Y /f_Y are called marginal PMFs/PDFs and are related to p_{XY} / f_{XY}:

- $\mathbf{p_X(k) = \sum p_{XY}(k,r)}$ **for all r** $\quad$ $\mathbf{f_X(x) = \int p_{XY}(x,y)dy}$ **for y $-\infty$ to ∞**
- $\mathbf{p_y(r) = \sum p_{XY}(k,r)}$ **for all k** $\quad$ $\mathbf{f_Y(y) = \int p_{XY}(x,y)dx}$ **for x $-\infty$ to ∞**

Bivariate normal example



$\mu_x = 1$

$\mu_y = 3$

$$\text{cov} = \begin{vmatrix} 5 & 0.3 \\ 0.3 & 2 \end{vmatrix}$$

$\rho = cov/(\sigma_x * \sigma_y)$

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

Equation taken from Wikipedia: https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Problem Tossing a coin: $\Omega = \{\mathbb{H}, \mathbb{T}\}$ and rolling strange die $\Omega = \{1, 2, 3\}$

If $\mathbb{H}$ we roll the die twice, if $\mathbb{T}$ we roll the die once

- Calculate joint PMF
- Calculate marginal PMF based on joint ones.

- (*) Calculate joint and marginal CDF

Problem Write a program that estimates the PMFs distribution

Problem Tossing a coin: $\Omega = \{\mathbb{H}, \mathbb{T}\}$ and rolling strange die $\Omega = \{1, 2, 3\}$

If $\mathbb{H}$ we roll the die twice, if $\mathbb{T}$ we roll the die once

X number of heads X = {0,1}
Y sum from die Y = {1,2,3,4,5,6}

- Calculate joint PMF
- Calculate marginal PMF based on joint ones.

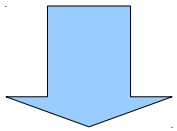
- (*) Calculate joint and marginal CDF

Conditional distribution

$A, B \in F$ such that $P(B) \neq 0$

Probability that event A happens “after” event B occurred:

$$\mathbf{P(A|B) = P(A \cap B)/P(B)}$$

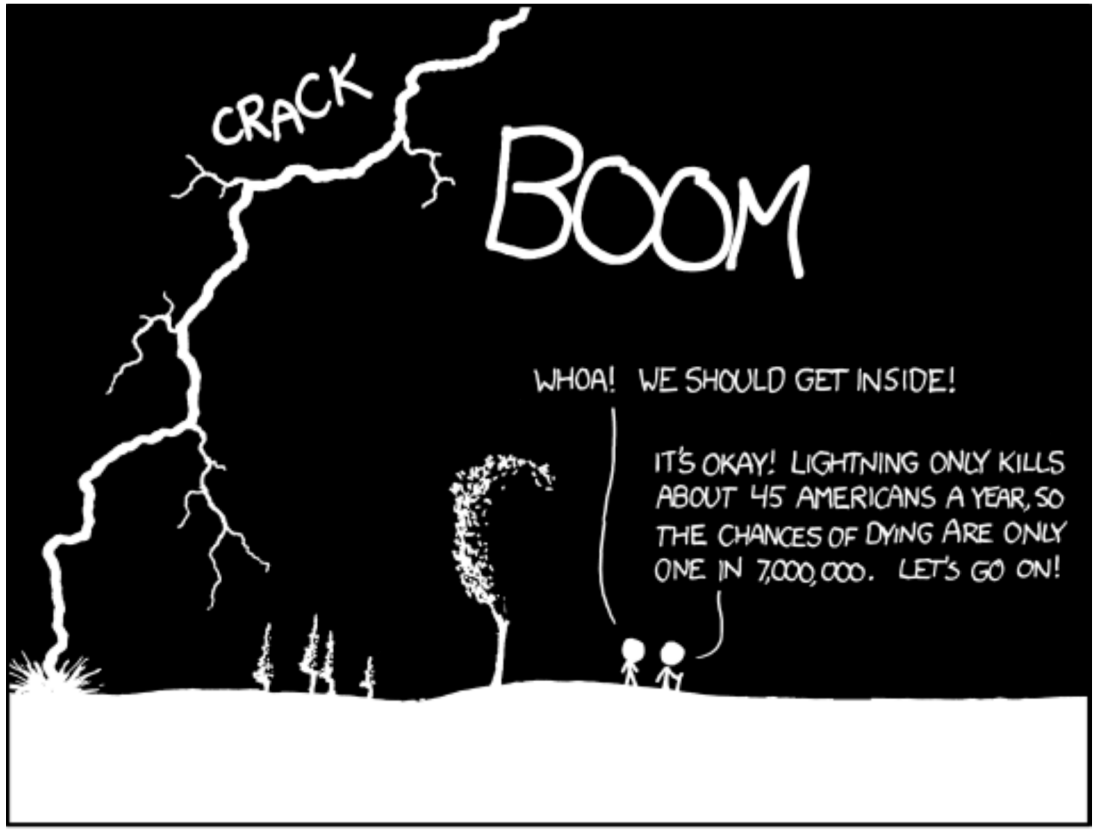


Conditional distribution

$A, B \in F$ such that $P(B) \neq 0$

Probability that event A happens “after” event B occurred:

$$P(A|B) = P(A \cap B)/P(B)$$



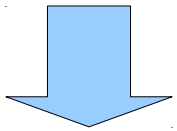
https://xkcd.com/795/

Conditional distribution

$A, B \in F$ such that $P(B) \neq 0$

Probability that event A happens “after” event B occurred:

$$\mathbf{P(A|B) = P(A \cap B)/P(B)}$$



X, Y two random variables

What is the probability distribution for Y when X random variable has value set to X=x ?

•(discrete) with the joint PMF: $p_{XY}(k,r)$, and $p_X(k) \neq 0$:

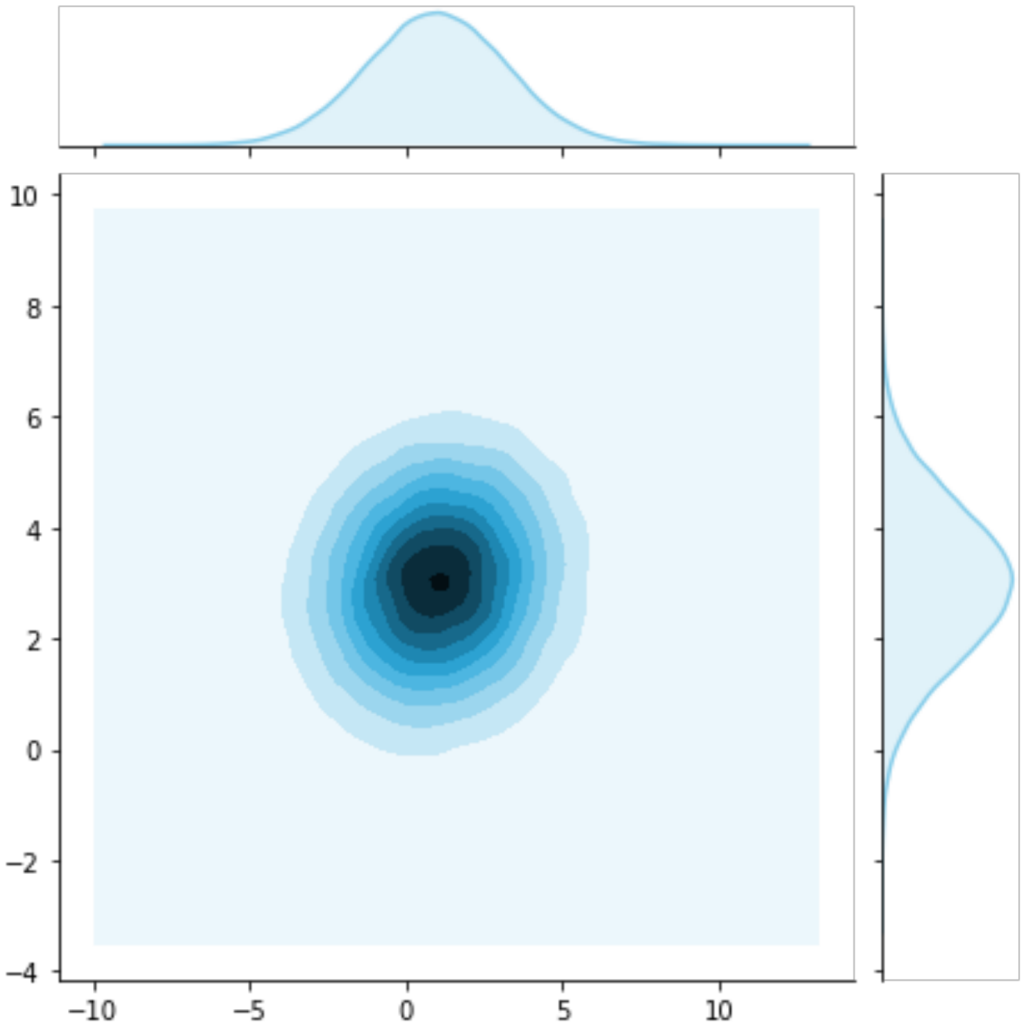
$$\mathbf{p_{Y|X}(r|k) = p_{XY}(k,r)/p_X(k)}$$

•(continuous) with the joint PDF $f_{XY}(x,y)$ and $f_X(x) \neq 0$

$$\mathbf{f_{Y|X}(y|x) = f_{XY}(x,y)/f_X(x)} \quad (*)$$

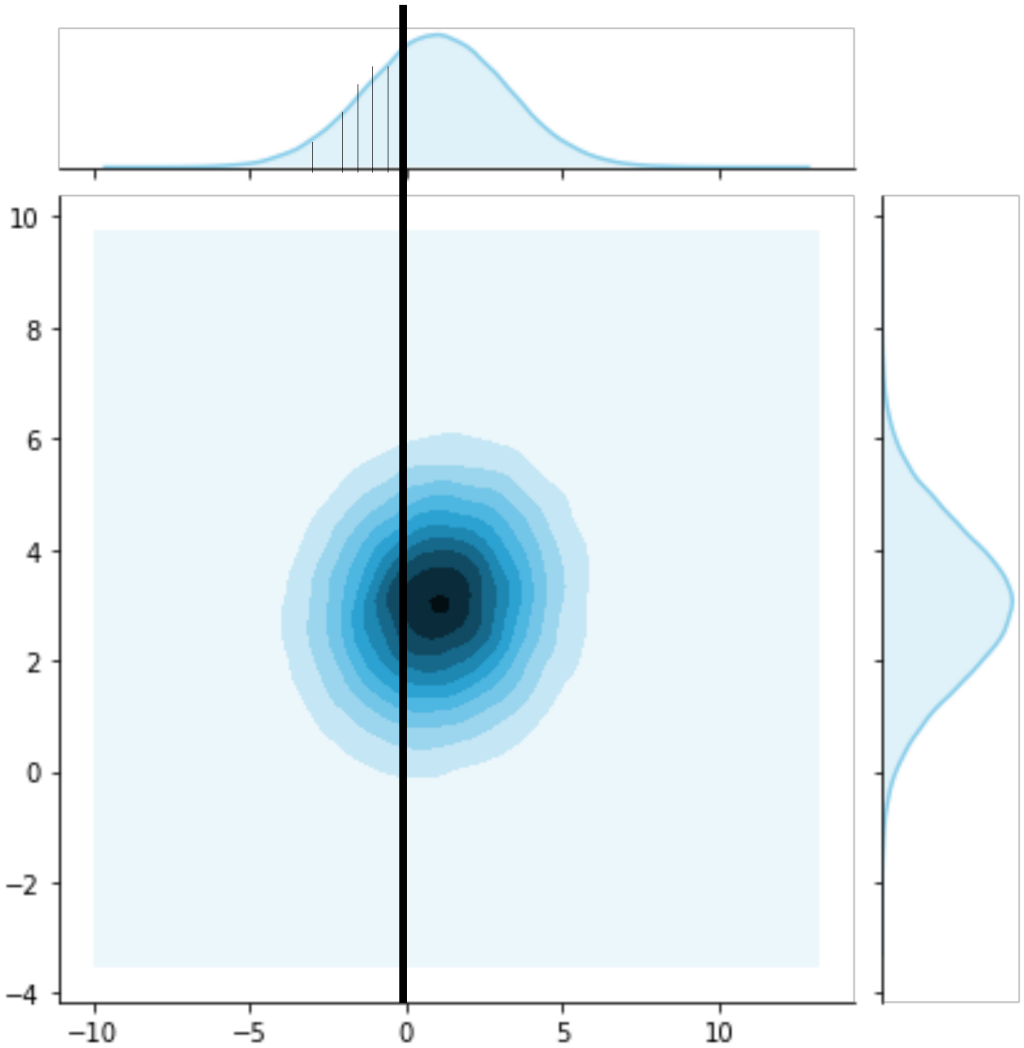
***Remark:** Formally in continuous case it should be derived using CDF and limit

Conditional probability



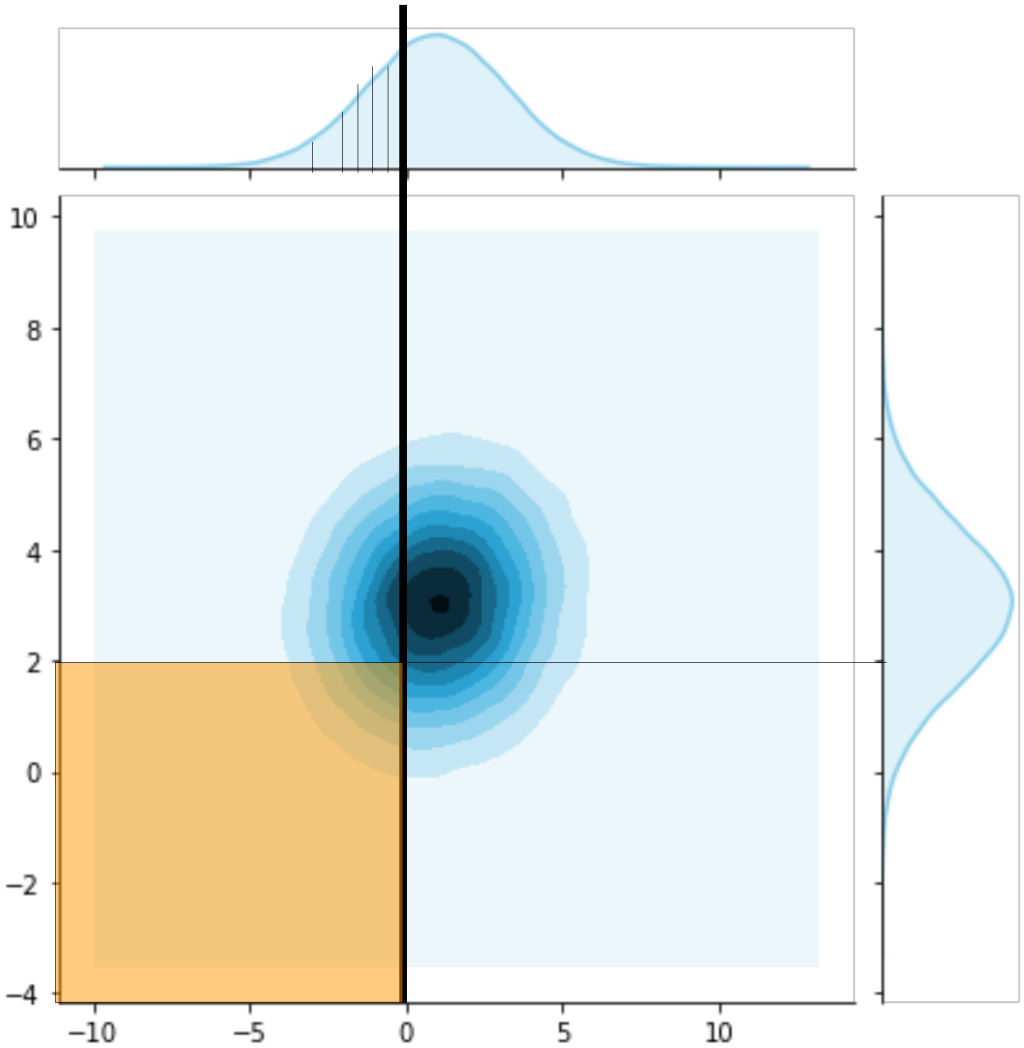
P(Y<=2 | X <=0) = P(X <=0 , Y <=2)/ P(X <=0)

Conditional probability



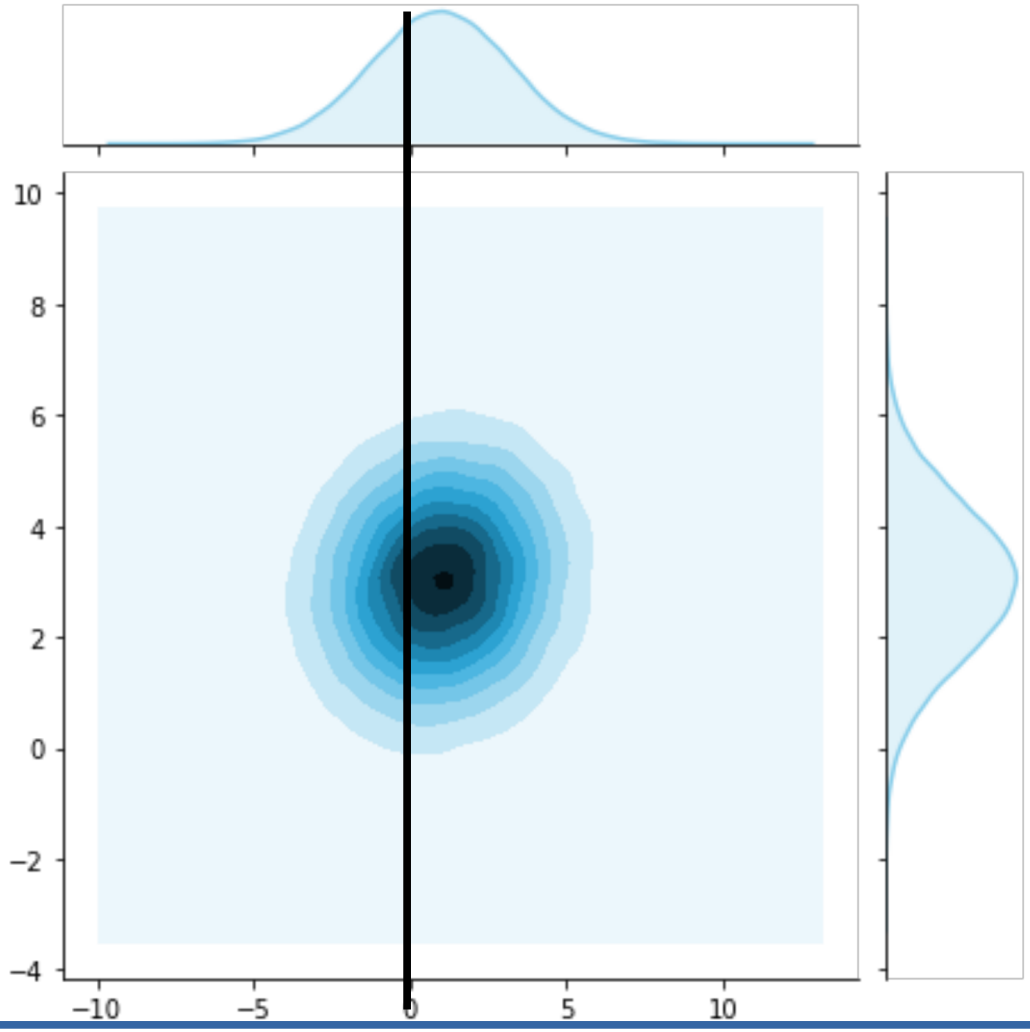
P(Y<=2 | X <=0) = P(X <=0 , Y <=2)/ P(X <=0)

Conditional probability



P(Y<=2 | X <=0) = P(X <=0 , Y <=2) / P(X <=0)

Conditional pdf



$$f_{Y|X}(y|\, x=0) = f_{XY}(x=0,y)/f_X(x=0)$$

- If you think all this is trivial – look for Borel-Kolmogorov paradox

Probability function vs likelihood function

X, Y two random variables

If we treat conditional PDF as a function of y: $g(y) = f_{Y|X}(y|x)$, it is a proper **probability function**:

$$\int g(y)dy = 1$$

If we treat conditional PDF as a function of x: $h(x) = f_{Y|X}(y|x)$, it is a **likelihood function:**

$$\int h(x)dx \text{ not necessarily } = 1$$

The same remark concerns PMFs

Problem: Rolling a die $\Omega = \{1, 2, 3, 4, 5, 6\}$

X is 1 if even number 0 otherwise.
Y is 1 if prime number 0 otherwise.

- Calculate joint PMF
- Calculate marginal PMF of X and of Y
- Calculate conditional PMF $p_{Y|X}(r|X=1)$
- Check if $h(k) = p_{Y|X}(r|k)$ is a proper probability function with respect to k

Problem: Let X and Y have a joint PDF $f_{XY}(x,y) = x+y$ for $0 < x < 1, 0 < y < 1$

- Find conditional PDF $f_{Y|X}(y|x)$
- Show that the integral of $f_{Y|X}(y|x)$ over all y values is equal to 1

Expectation for X and Y

Let g(x,y) be some function: $R^2 \rightarrow R$, X, Y – random variables, g(X,Y) becomes random variable as well:

- For X,Y being discrete random variables with joint PMF $p_{XY}(k,r)$:

  $E[g(X,Y)] = \Sigma\Sigma\ g(k,r)\ p_{XY}(k,r)$ for all possible values of k and r

- For X,Y being a continuous random variables with PDF $f_X(x,y)$:

  $E[g(X,Y)] = \int\int g(x,y)\ f_{XY}(x,y)\ dxdy$ for x $-\infty$ to ∞, y $-\infty$ to ∞

Properties:

- **E[f(X,Y) +g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]** - linearity
- E[K* f(X,Y)] = K *E[f(X,Y)] for K being constant

- **Intuition:** E[g(X,Y)] is a kind of weighted average with weights defined by p_{XY} or f_{XY}

Covariance

Covariance expresses the relationship between two random variables X,Y

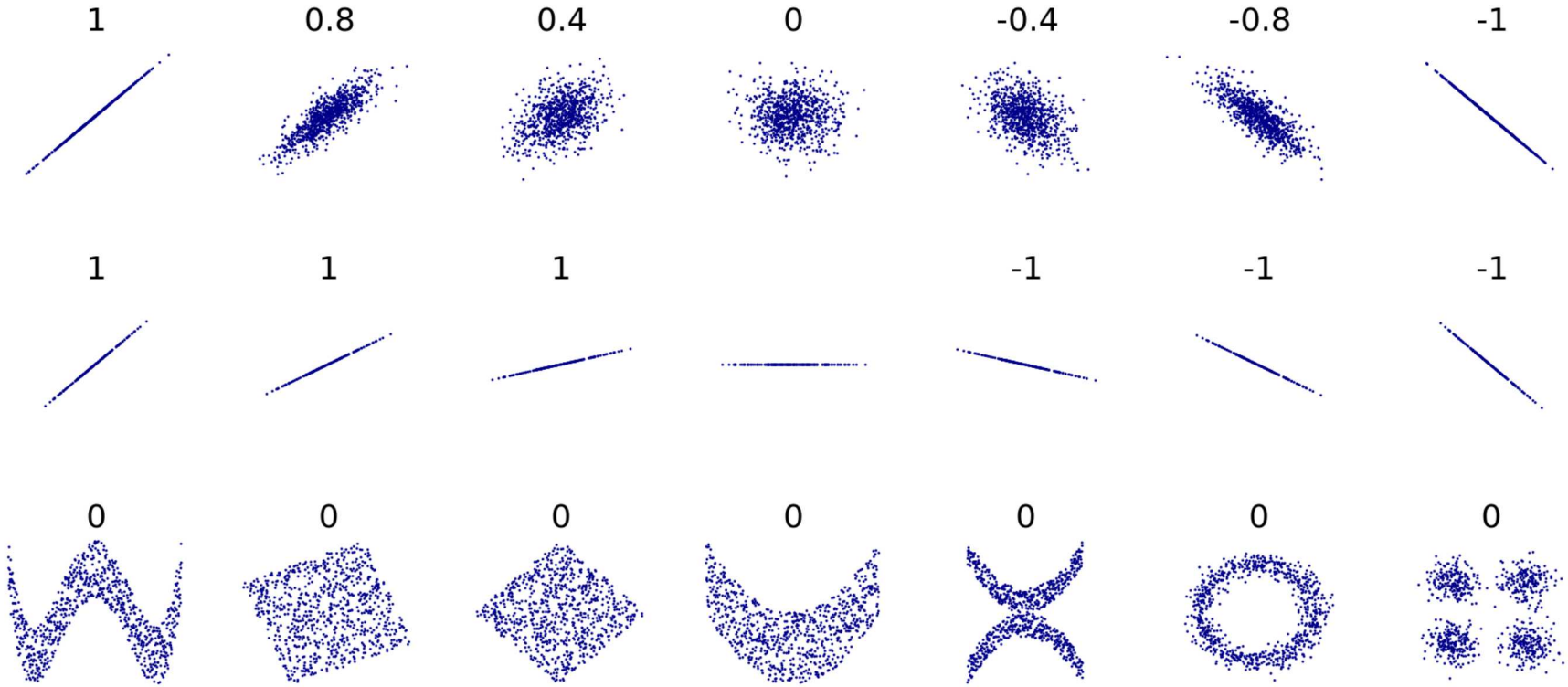
$Cov[X,Y] = E[(X - E[X])*(Y - E[Y])] \rightarrow Cov[X,Y] = E[XY] - E[X]E[Y]$

- Covariance captures only **linear** dependency between variables (!!):
  - If Cov[X,Y] >0 then if X increases then Y will increase,
  - If Cov[X,Y] <0 then if X increases then Y will decrease,
  - If Cov[X,Y] = 0 we say that X and Y are **uncorrelated.**

- Cov[X,X] = Var[X]
- Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]

Remark: If X,Y are independent, then Cov[X,Y] = 0 but opposite not always true

Covariance



From Wikipedia by Denis Boigelot, https://commons.wikimedia.org/w/index.php?curid=15165296

Independence

A, B ∈ F are independent if:

$P(A \cap B) = P(A) * P(B) \leftrightarrow P(A|B) = P(A), P(B|A) = P(B)$

We want to generalize the notion of independence to random distributions.

Let X,Y be two random variables.

Independent variables X, Y

Let X, Y be **independent** random variables, then it is true that:

- E[f(X) *g(Y)] = E[f(X)] * E[g(Y)] → factorization of expectation values

Independent variables X, Y

Let X, Y be **independent** random variables, then it is true that:

- $E[f(X) * g(Y)] = E[f(X)] * E[g(Y)]$ → factorization of expectation values
- $F_{XY}(x,y) = F_X(x) * F_Y(y)$ → factorization of joint CDF

Independent variables X, Y

Let X, Y be **independent** random variables, then it is true that:

- $E[f(X) * g(Y)] = E[f(X)] * E[g(Y)]$ → factorization of expectation values
- $F_{XY}(x,y) = F_X(x) * F_Y(y)$ → factorization of joint CDF
- For X, Y, being discrete random variables:
  - $p_{XY}(k,r) = p_X(k) * p_Y(r)$ for all k from X and r from Y → factorization of PMF
  - $p_{Y|X}(r|k) = p_Y(r)$ for all r from Y, where $p_X(k) \neq 0$

Independent variables X, Y

Let X, Y be **independent** random variables, then it is true that:

- $E[f(X) * g(Y)] = E[f(X)] * E[g(Y)]$ → factorization of expectation values
- $F_{XY}(x,y) = F_X(x) * F_Y(y)$ → factorization of joint CDF
- For X, Y, being discrete random variables:
  - $p_{XY}(k,r) = p_X(k) * p_Y(r)$ for all k from X and r from Y → factorization of PMF
  - $p_{Y|X}(r|k) = p_Y(r)$ for all r from Y, where $p_X(k) \neq 0$
- For X, Y, being continuous random variables:
  - $f_{XY}(x,y) = f_x(x) * f_Y(y)$ for all x,y from R → factorization of PDF
  - $f_{Y|X}(y|x) = f_Y(y)$ for all y from Y, where $f_X(x) \neq 0$

Problem: Show that:

- $Cov[X,Y] = E[XY] - E[X]E[Y]$
- $Var[X+Y] = Var[X] + Var[Y] + 2\,Cov[X,Y]$

Problem: Let X be uniform in (-1,1) and $Y = X^2$

- Check if X and Y are correlated
- Check if X and Y are independent

Conditional expectation

Conditional expectation of random variable X expresses E[X] under some set of conditions that “occurred”.

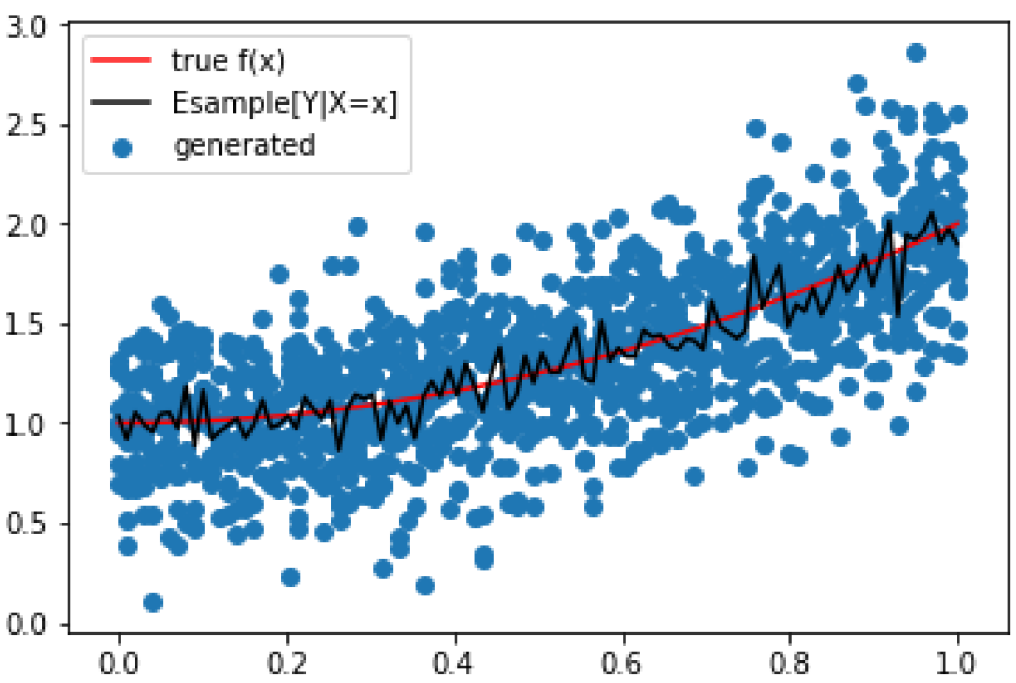
- For X,Y being discrete random variables with conditional PMF $p_{Y|X}(r|k)$:

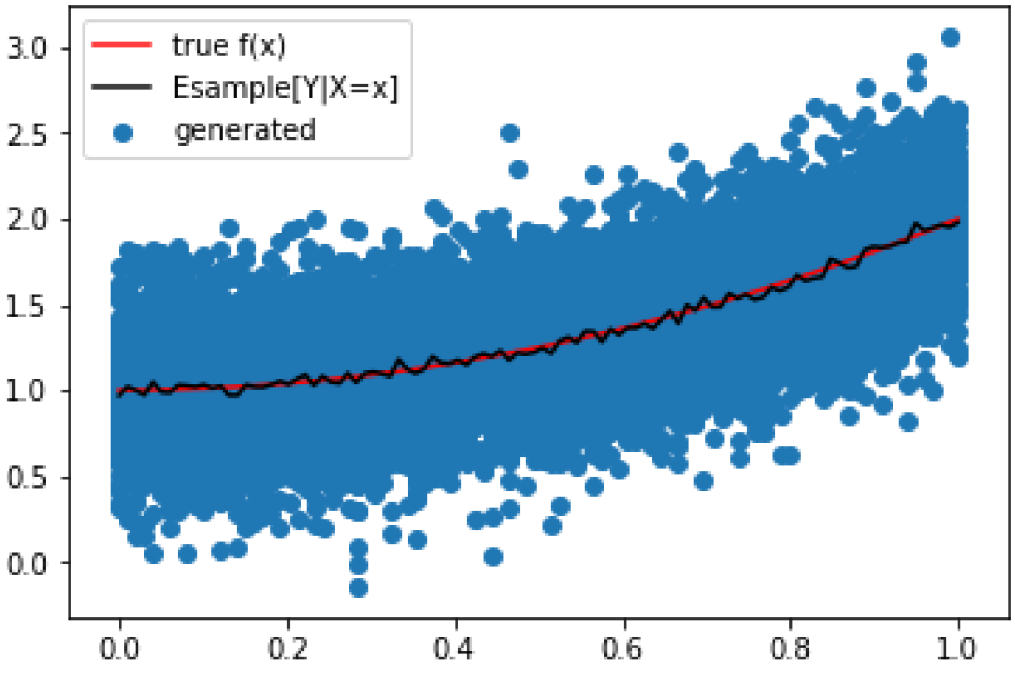
  $E[Y|k] = \sum r * p_{Y|X}(r|k)$ for all possible values of r

- For X,Y being a continuous random variables with conditional PDF $f_{Y|X}(y|x)$:

  $E[Y|x] = \int y * f_{Y|X}(y|x)\, dy$ for y from $-\infty$ to ∞

Conditional expectation example





Conditional variance

Conditional variance of random variable X expresses Var[X] under some set of conditions that “occurred”.

- For X,Y being discrete random variables with conditional PMF $p_{Y|X}(r|k)$:

  $Var[Y|k] = \sum (r - E[Y|k])^2 * p_{Y|X}(r|k)$ for all possible values of r

- For X,Y being a continuous random variables with conditional PDF $f_{Y|X}(y|x)$:

  $Var[Y|x] = \int (y - E[Y|x])^2 * f_{Y|X}(y|x)\, dy$ for y from $-\infty$ to ∞

- One can show that :

  $Var[Y|k] = E[Y^2|k] - (E[Y|k])^2$

Problem 6 rolling a die $\Omega = \{1, 2, 3, 4, 5, 6\}$

X is 1 if even number 0 otherwise.
Y is 1 if prime number 0 otherwise.

- Calculate E[Y|1], E[Y|0]
- Var[Y|1], Var[Y|0]

Problem:

- Implement a function that returns mean
- Implement a function that returns Var[X]
- Implement a function that returns Cov[X,Y]

Problem: Similarity measures

- Implement a function that returns Euclidean distance between two vectors
- Implement a function that returns Manhattan distance between two vectors
- Implement a function that returns Cosine similarity
- Implement a function that returns Pearson correlation coefficient

Thank you