INTRODUCTION TO MACHINE LEARNING

DIRK HESSE

STK-INF4000, WEEK 4

MACHINE LEARNING EXAMPLES

- Classify whether an email is spam or not, given the words in in it.
- Predict demand for a certain article, given past demand for similar articles.
  - How many copies of a book should we print?
  - How may gadgets do we need to stockpile?
- Predict CO_2 levels today, given yesterday's levels and weather data.
- Predict relevance of a search result for a given user.
- Predict the sentiment of a product review.

ROUGH DEFINITION

Machine learning makes use of computers to interpret data and find patterns in it, often enabling us to predict properties of instances of the data not yet seen.

DATA

- We will denote our inputs X, and our target Y.
  - *Example*: X contains CO2 levels, temperature, etc. on a given day, the CO2 levels the day after.
- X is often a vector, $X^T = (X_1, \ldots, X_p)$.
- Y is often a scalar, but could be a vector as well.

LEARNING TASK

Given examples for X and Y, denoted

$$(x_i, y_i) \quad i = 1, \ldots, n$$

find a function that gives a reasonable (for some value of reasonable) prediction $\hat{y}$ given a previously unseen sample x.

$x_i = \left(x_i^{(1)}, \ldots, x_i^{(p)}\right)^T$ and y_i are properties of the i-th example, e.g. time spent on a website and numbers of links clicked during a visit of a specific user.

TYPES OF DATA

- Continuous
  - E.g. height, CO_2 concentration.
- Categorical
  - Spam (yes or no), color.
- Ordered categorical
  - High, medium, low.

TYPES OF LEARNING

- Supervised learning.
  - As explained.
  - Given (x_i, y_i), find a function to predict y given x.
  - Examples: Spam classification, demand prediction.
- Unsupervised learning.
  - No target Y.
  - Find patterns in the data.
  - Examples: Recommender systems, finding groups.

WHAT CAN BE LEARNED?

- Y *must* be dependent on X.
- Dependence can be *very* complex.
  - Example: Sentiment analysis.
  - Deep learning.

SOME BASIC PROBABILITY

PROBABILITY DISTRIBUTIONS

- Discrete case:
  - Probability mass function, PMF: $\mathrm{P}[X]$
  - Example: $P[heads] = 0.5$.
- Continuous case:
  - Probability density function, PDF: $p(x)$

EXPECTATION VALUES

DISCRETE CASE

$$\mathrm{E}[X] = \sum_x x \; \mathrm{P}[X = x]$$

$$\mathrm{E}[f(X)] = \sum_x f(x) \; \mathrm{P}[X = x] \neq f(\mathrm{E}[X])$$

EXPECTATION VALUES

CONTINUOUS CASE

$$\mathrm{E}[X] = \int x\, p(x) \mathrm{d}x$$

$$\mathrm{E}[f(X)] = \int f(x)\, p(x) \mathrm{d}x \neq f(\mathrm{E}[X])$$

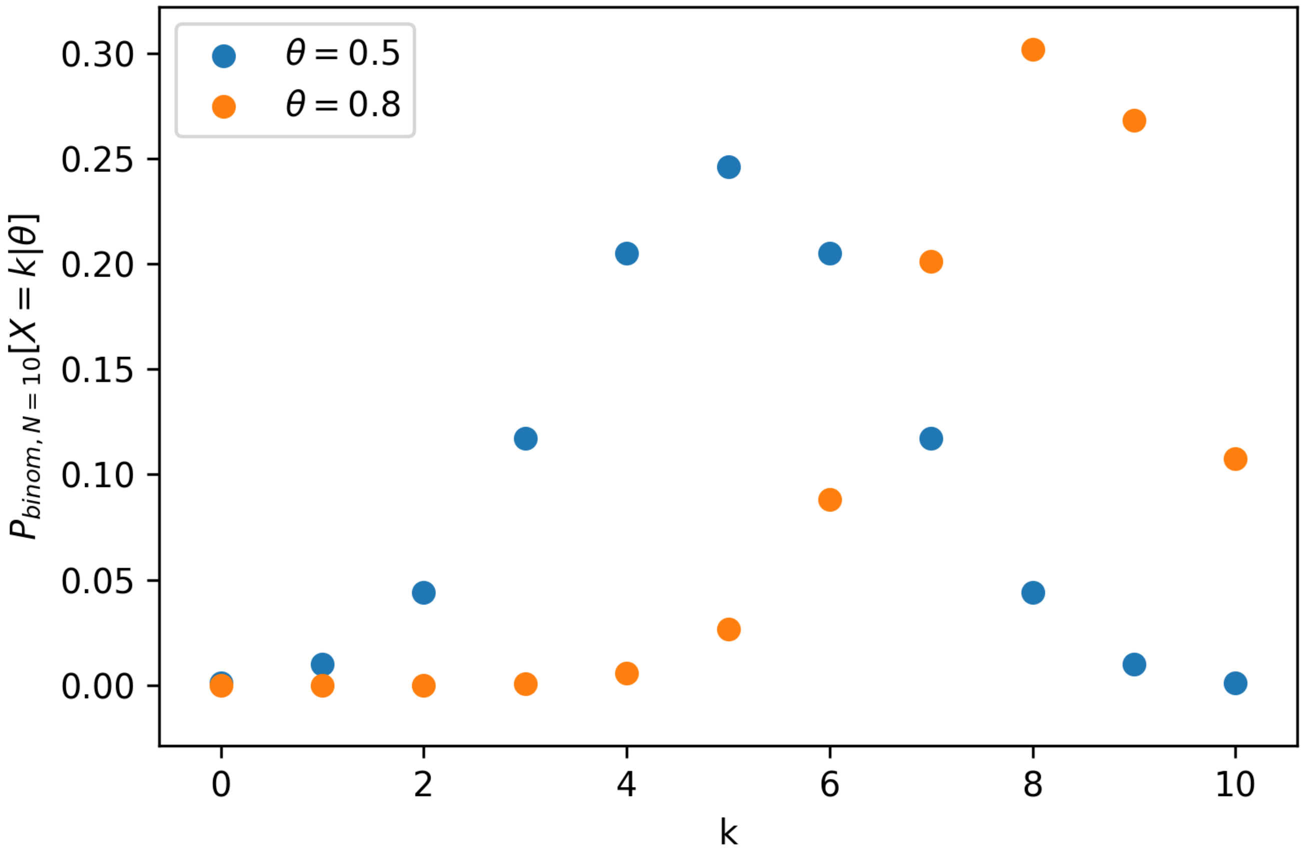
FAMOUS PROBABILITY DISTRIBUTIONS

DISCRETE DISTRIBUTIONS

BINOMIAL

$$\mathrm{P}[X = k|\theta] = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

- Email is spam with probability θ, then $\mathrm{P}[X = k|\theta]$ is the probability of having k out of N emails spam.
- Production errors.
- Click rate.

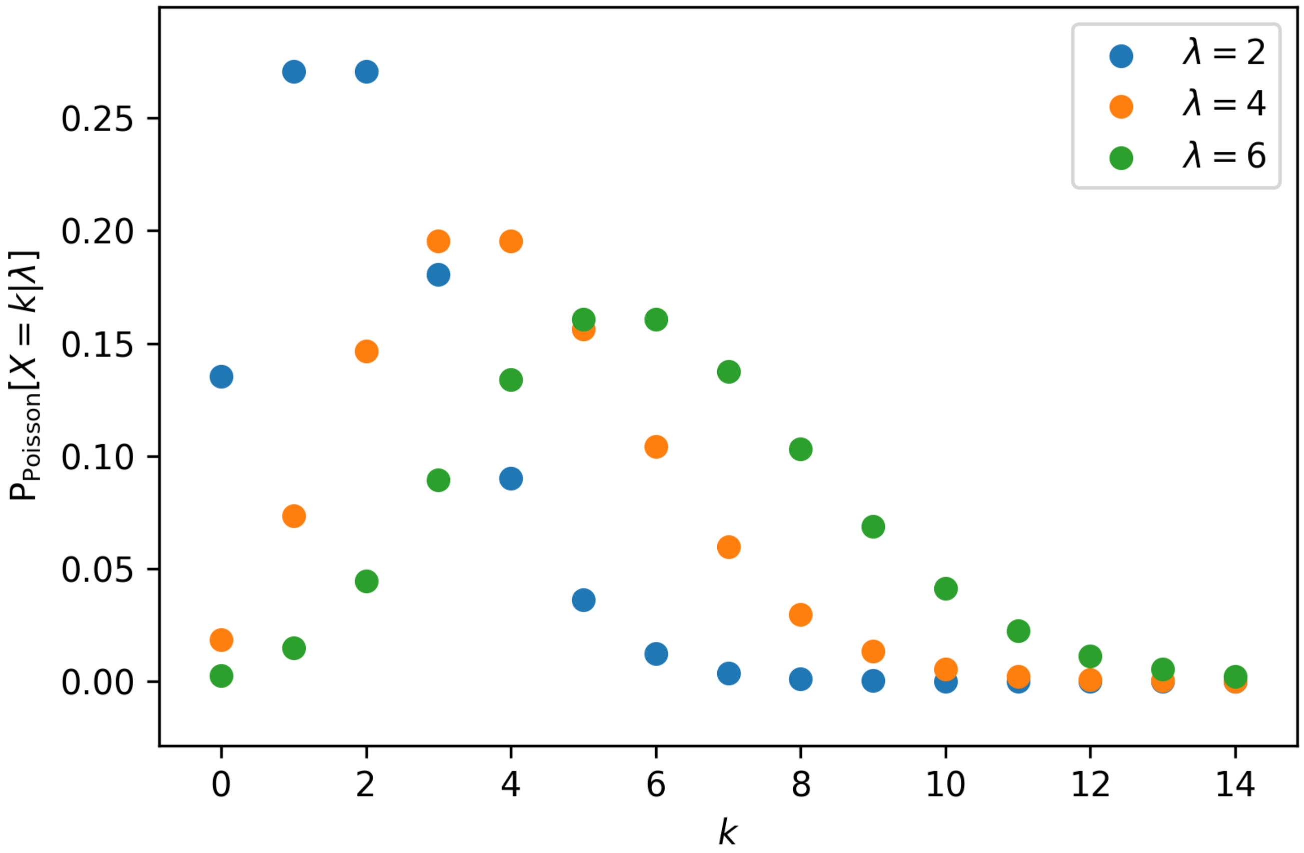


$\theta = 0.5$
$\theta = 0.8$
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$P_{binom, N=10}[X = k|\theta]$
0
2
4
6
8
10
k

POISSON

$$\mathrm{P}[X = k|\lambda] = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Events occurring at a rate λ.
- Radioactive decays.
- Cars arriving at intersection.
- Number of network packets arriving.



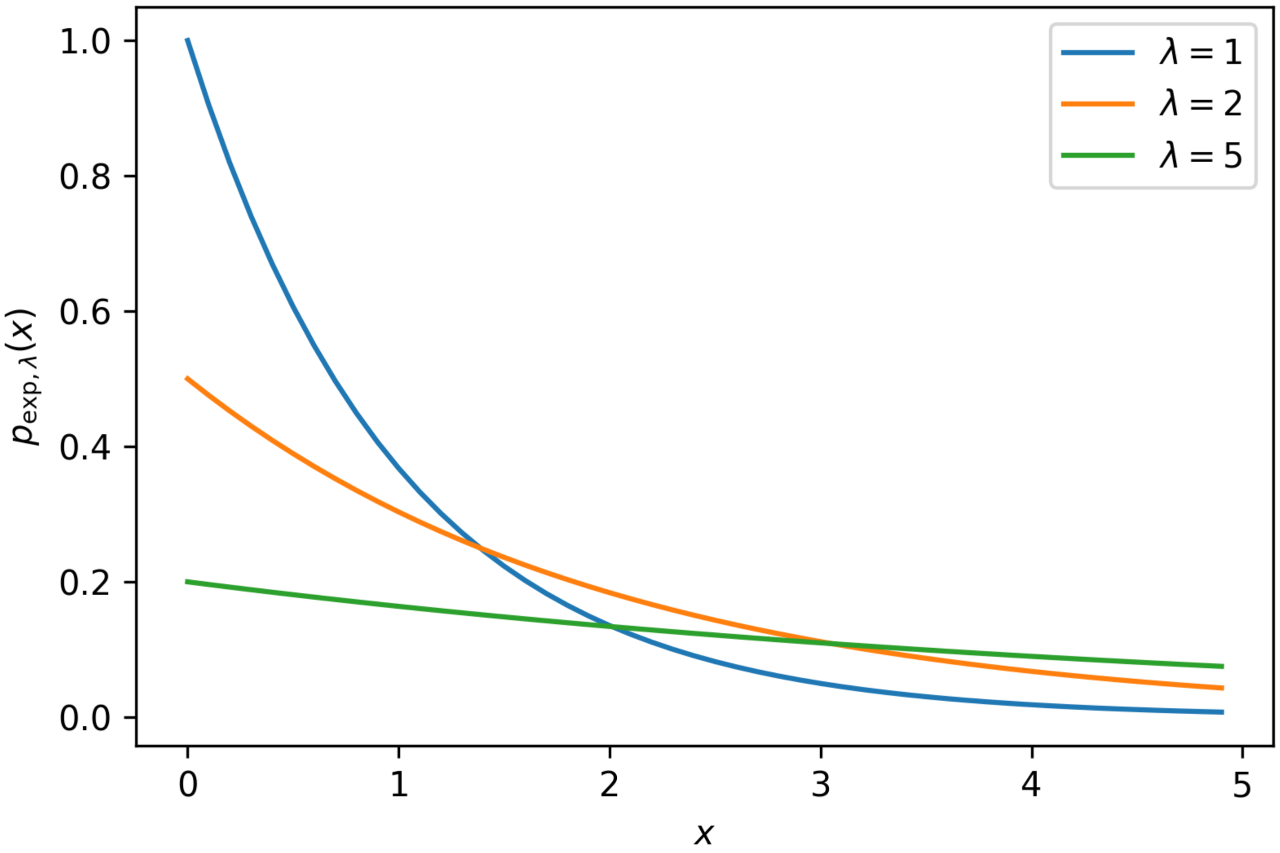
$\lambda = 2$
$\lambda = 4$
$\lambda = 6$
0.25
0.20
0.15
0.10
0.05
0.00
$P_{\text{Poisson}}[X = k|\lambda]$
0
2
4
6
8
10
12
14
k

CONTINUOUS DISTRIBUTIONS

EXPONENTIAL

$$p(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, \quad x > 0$$

- Web server response time.
- Time to next earthquake.
- Time until hard drive failure.

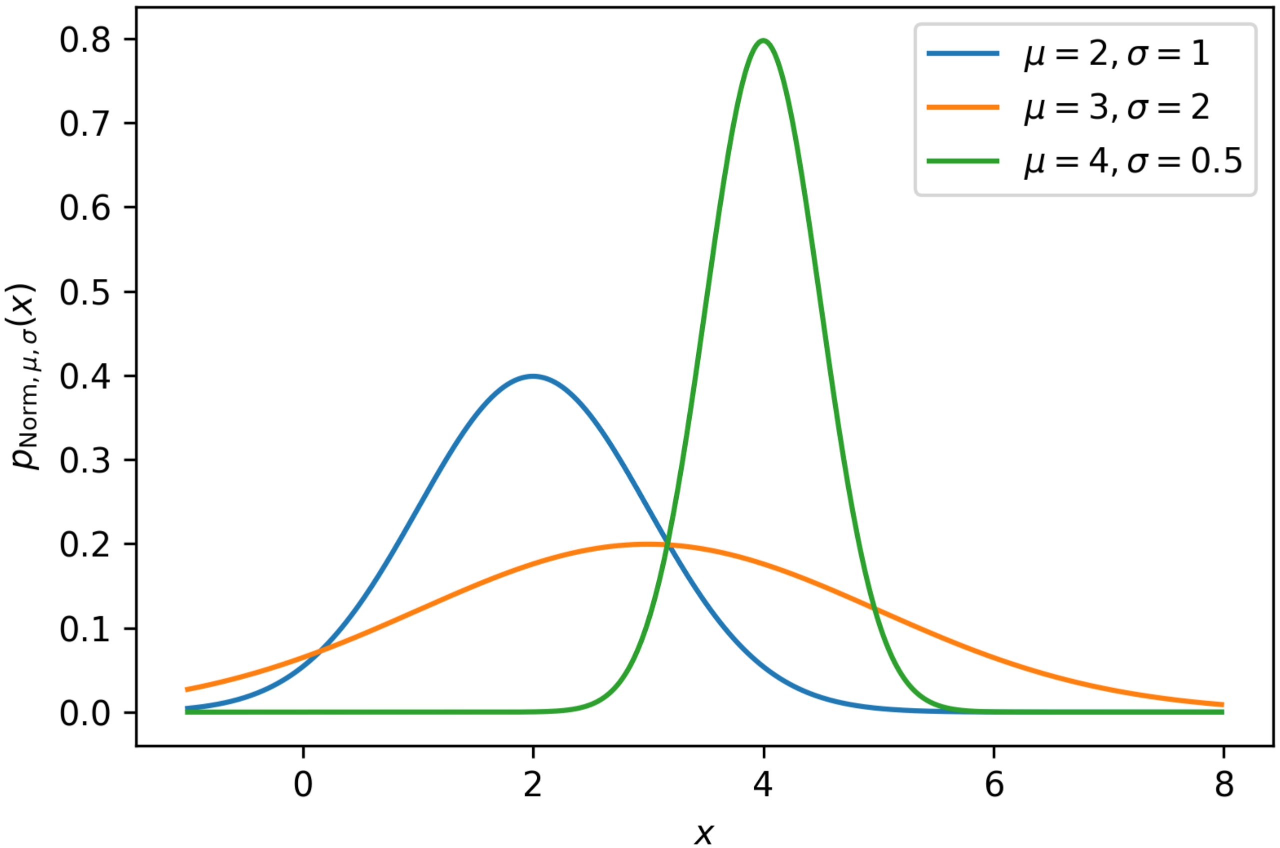


$\lambda = 1$
$\lambda = 2$
$\lambda = 5$
1.0
0.8
0.6
0.4
0.2
0.0
$p_{\exp,\lambda}(x)$
0
1
2
3
4
5
x

GAUSSIAN / NORMAL

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- CO2 levels.
- Telnet session length.
- Student scores.



0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
$p_{\mathrm{Norm},\mu,\sigma}(x)$
0
2
4
6
8
x
$\mu = 2, \sigma = 1$
$\mu = 3, \sigma = 2$
$\mu = 4, \sigma = 0.5$

CONDITIONAL PROBABILITY

Example: Study concerning income class (high, low).

	Female	Male	
High	9%	11%	20%
Low	46%	34%	80%
	55%	45%	

- 9%, 11%, ...: joint probabilities.
- 20%, 60%, ...: marginal probabilities.

CONDITIONAL PROBABILITY

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

Here, $P(High|Female) \approx 16\%$, while $P(High|Male) \approx 24\%$.

WHAT DOES THIS HAVE TO DO WITH ML?

DECISION THEORY - SUPERVISED LEARNING

We have

- $X \in \mathbb{R}^p$ random input vector.
- $Y \in \mathbb{R}$ output vector.
- Looking for $f(X)$ predicting Y.

WHY?

PREDICTION

- Exact form of f not too important.
- Accuracy important.

INFERENCE

- Which parts of X are important for predicting Y?
- What is the relationship between Y and X?
- What is the *functional* relationship between them? Linear?
- **Exact form** of f *is* important.

HOW TO GET f?

- *Loss function* $L(Y, f(X))$,
  - Often $L(Y, F(X)) = (Y - f(X))^2$.
- Minimize expected prediction error.

$$\mathrm{EPE}(f) = E(Y - f(x))^2 = \int (y - f(x))^2 p(x, y) dx\, dy$$

CAREFUL, MATH!

Remember $\mathrm{P}(X,Y) = \mathrm{P}(Y|X)\,\mathrm{P}(X)$? This gives

$$\mathrm{EPE}(f) = \mathrm{E}_X\,\mathrm{E}_{X|Y}\big([Y - f(X)]^2|X\big)$$

and thus we minimize point by point

$$f(x) = \operatorname*{argmin}_c\ \mathrm{E}_{Y|X}\big([Y - c]^2|X = x\big)$$

This yields

$$f(x) = \mathrm{E}(Y|X = x)$$

HOW GOOD CAN WE DO?

- Assume $Y = f(X) + \epsilon$.
  - $\epsilon \sim N(0, \sigma)$ (normal distributed).
- Our estimate $\hat{f}(X) = \hat{Y}$:

$$\mathrm{E}(Y - \hat{Y})^2 = \mathrm{E}[f(X) + \epsilon - \hat{f}(X)]^2 = \mathrm{E}[f(X) - \hat{f}(X)]^2 + \mathrm{Var}($$

- We have control over $\mathrm{E}[f(X) - \hat{f}(X)]^2$.
- We have no control over $\mathrm{Var}(\epsilon)$.

K-NEAREST NEIGHBORS

KNN implements this in a very simple way:

$$\hat{f}(x) = \frac{1}{k} \sum_{z \in N_k(x)} z \, ,$$

where $N_k(x)$ are the k closest training examples to x from a given trainir set.

WHAT ABOUT OTHER LOSS FUNCTIONS?

$$L(Y, f(X)) = |Y - f(X)|$$

leads to

$$f(x) = \mathrm{median}[Y|X = x]$$

Other choices are possible!