

Lecture 5

Корбут Даниил

Deep Learning Research Engineer, Insilico Medicine

telegram: @rtriangle

vk: rtriangle

email: korbut.daniel@gmail.com

План занятия

- Напоминание прошлой лекции
- Временные ряды и NN
- Рекомендательные системы и NN
- Кластеризация и методы понижения размерности
- Autoencoder
- Summary курса

Sequence labeling

Problem

- Given a sequence of *tokens*, infer the most probable sequence of *labels* for these tokens.

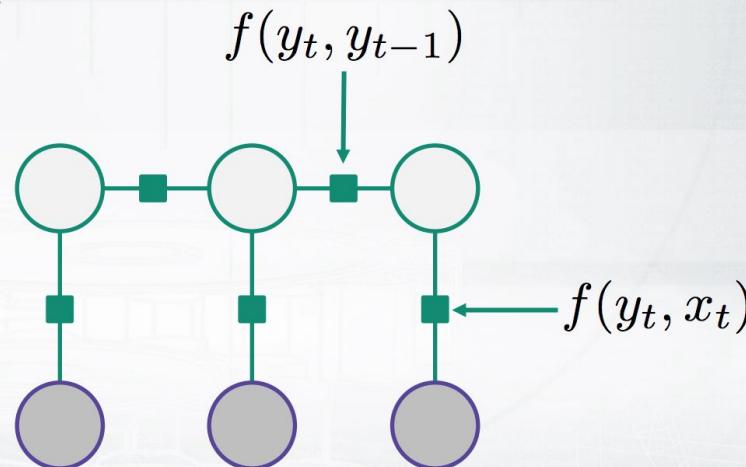
Examples

- part of speech tagging
- named entity recognition

Conditional Random Field (linear chain)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(x)} \prod_{t=1}^T \exp \left(\sum_{k=1}^K \theta_k f_k(y_t, y_{t-1}, x_t) \right)$$

Undirected graph:



Features engineering

Label-observation features:

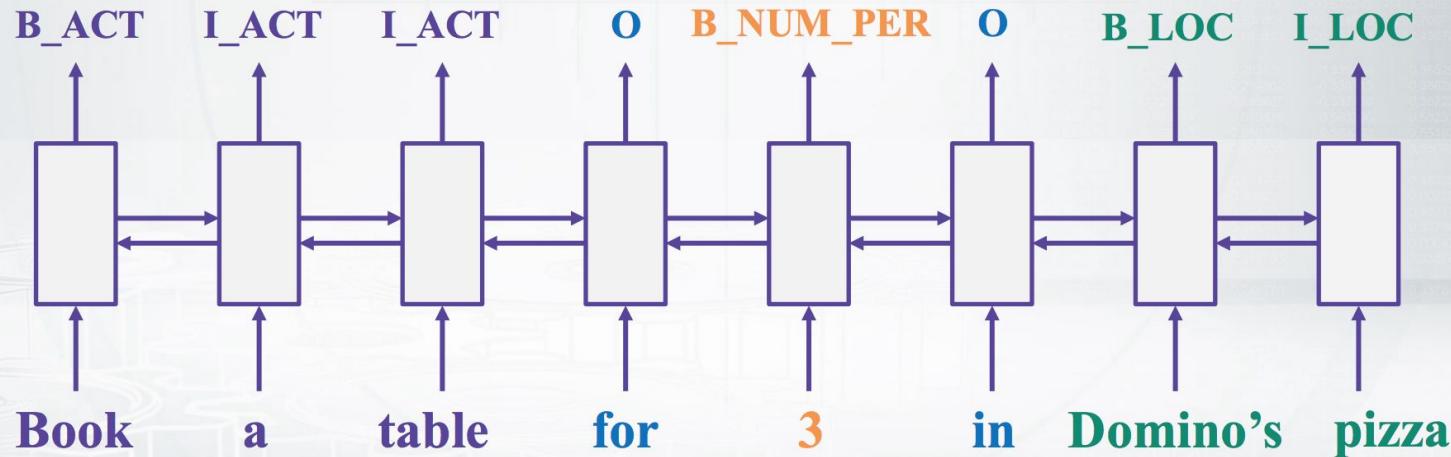
$$\bullet \quad f(y_t, y_{t-1}, x_t) = [y_t = y] g_m(x_t)$$

$$\bullet \quad f(y_t, y_{t-1}, x_t) = [y_t = y][y_{t-1} = y']$$

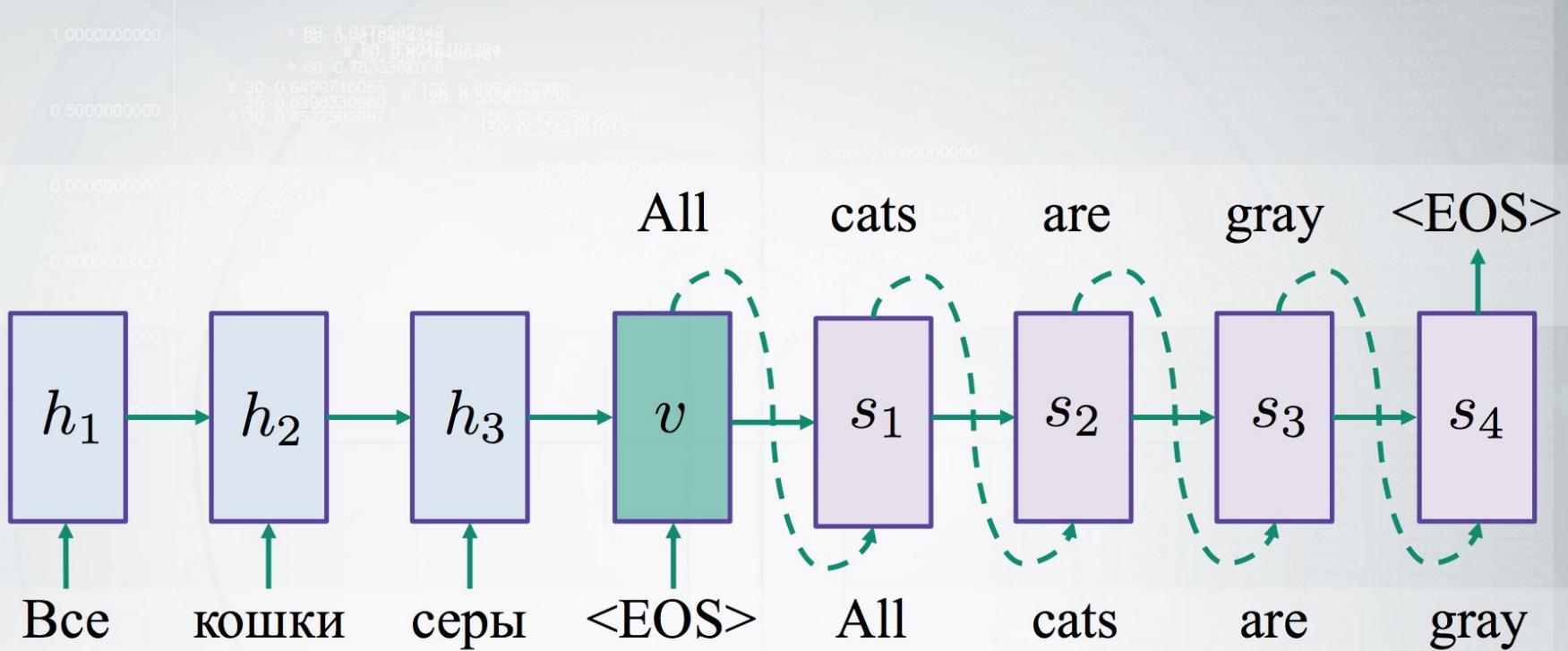
$$\bullet \quad f(y_t, y_{t-1}, x_t) = [y_t = y][y_{t-1} = y'] g_m(x_t)$$

Bi-directional LSTM

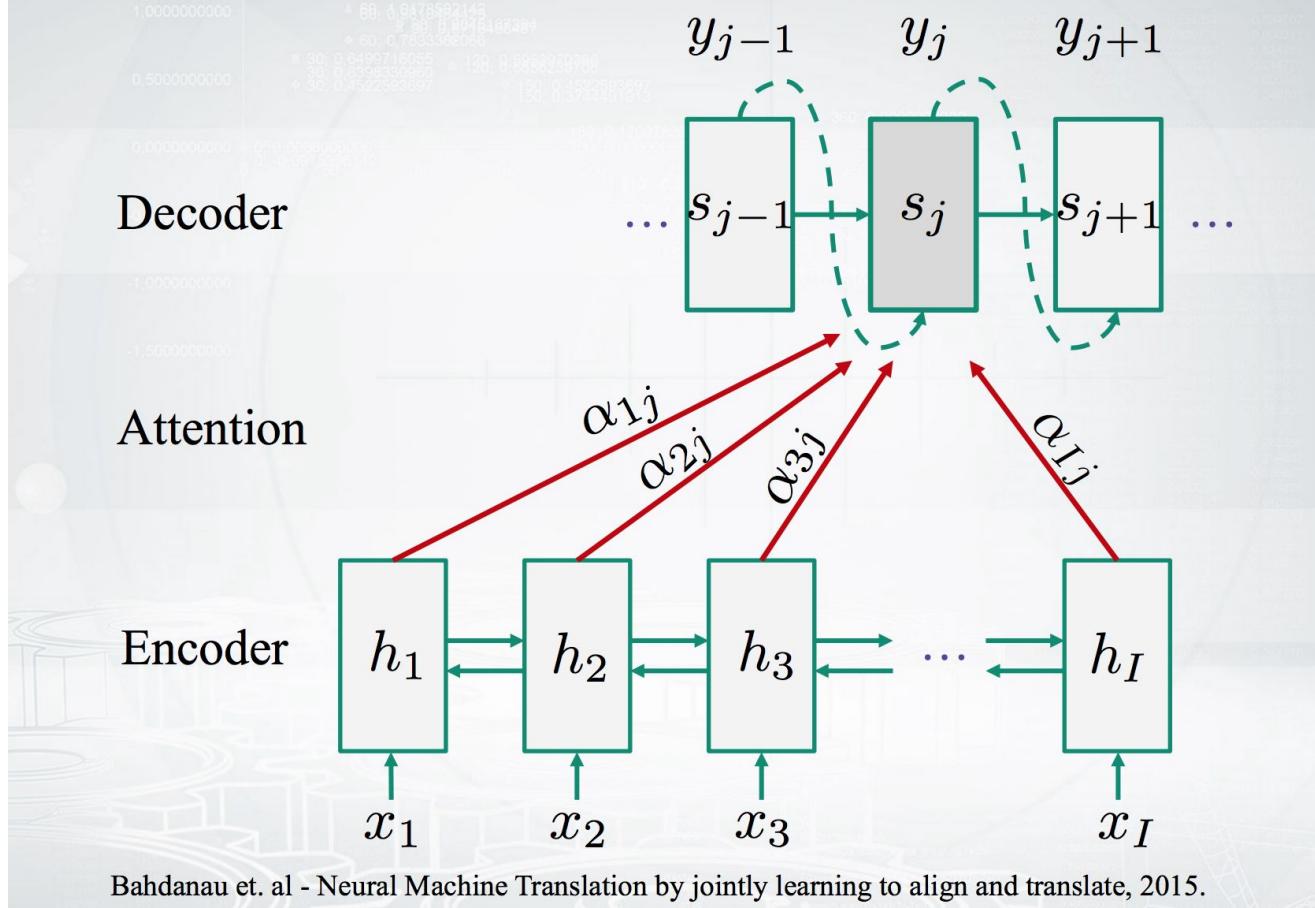
- Universal approach for sequence tagging
- You can stack several layers + add linear layers on top
- Trained by cross-entropy loss coming from each position



Sequence to sequence



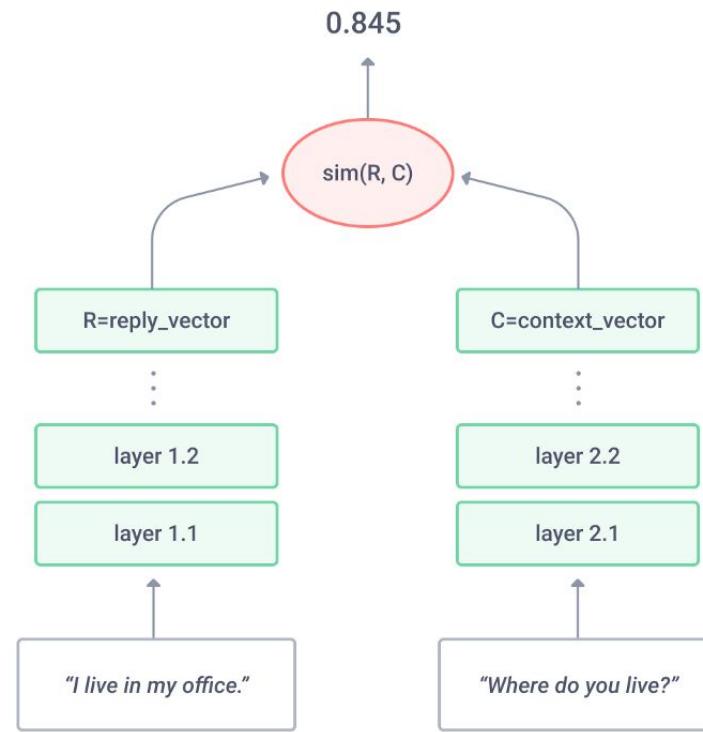
Attention mechanism



Bahdanau et. al - Neural Machine Translation by jointly learning to align and translate, 2015.

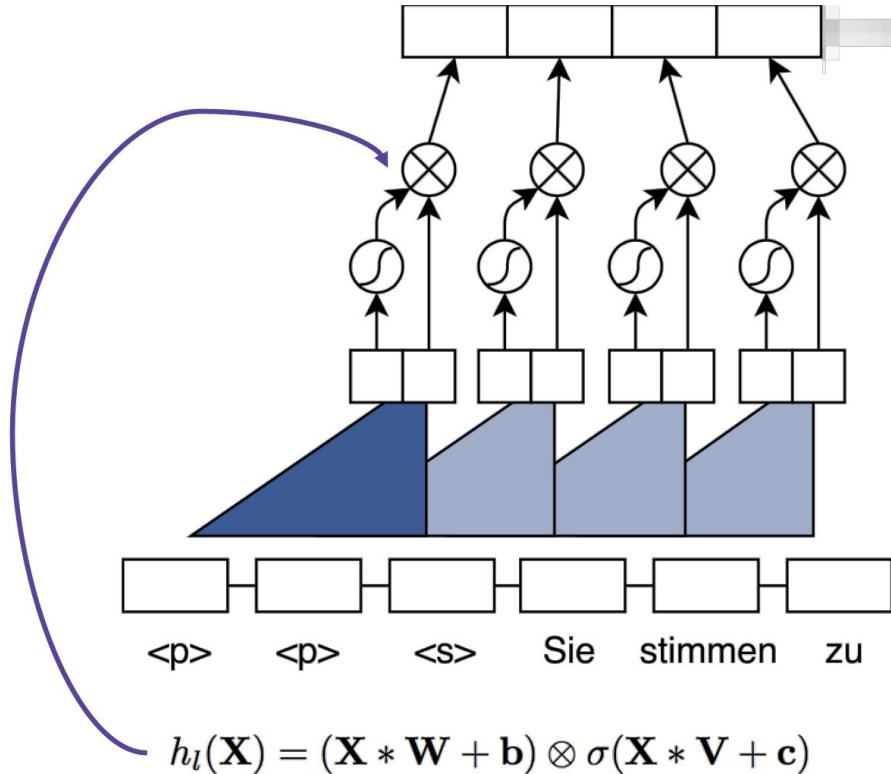
Selective models

Instead of estimating probability $p(\text{reply} | \text{context}; w)$, selective models learn similarity function— $\text{sim}(\text{reply}, \text{context}; w)$, where a reply is one of the elements in a predefined pool of possible answers.



$$L = \max(0, \text{sim}(\text{ctx}, \text{reply}_{\text{wrong}}) - \text{sim}(\text{ctx}, \text{reply}_{\text{correct}}) + \alpha) \rightarrow \min$$

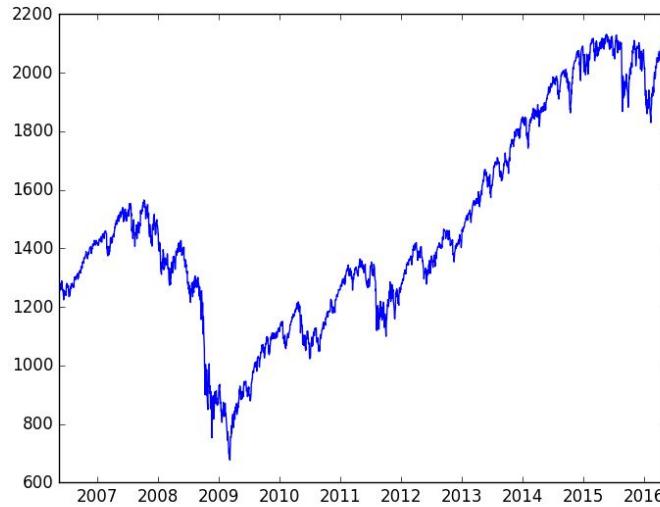
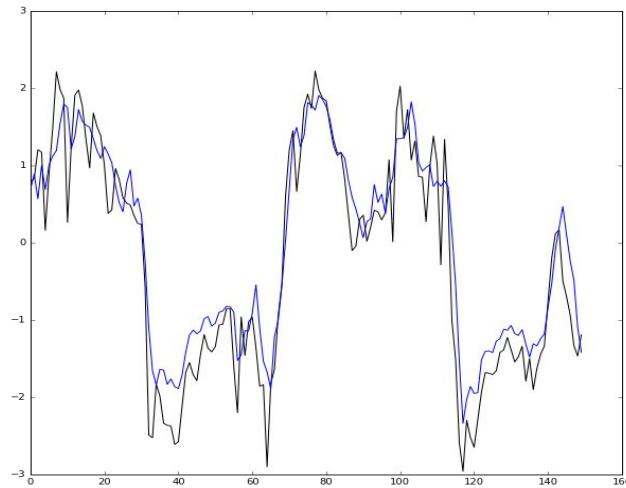
CNN for sequences: Gated Linear Unit



Stacking 6 layers
with kernel size 5
results in an input
field of 25 elements

Time series

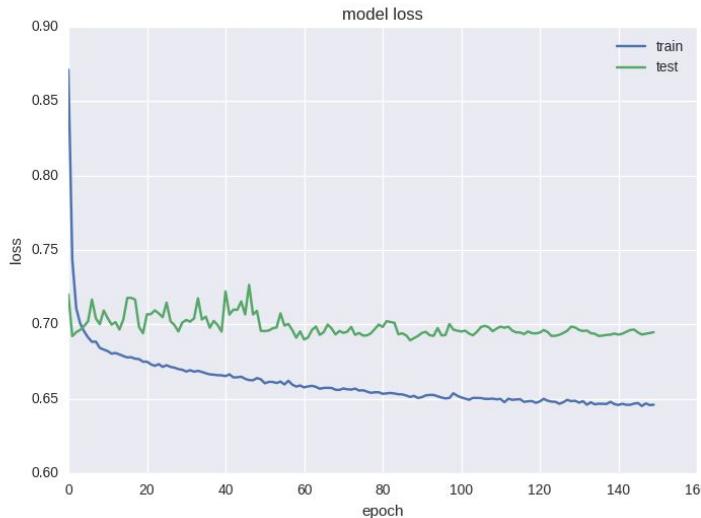
- Предсказание следующего значения: регрессия в случае точного показателя, классификация в случае определения направления изменения цены
- Поиск различных аномалий



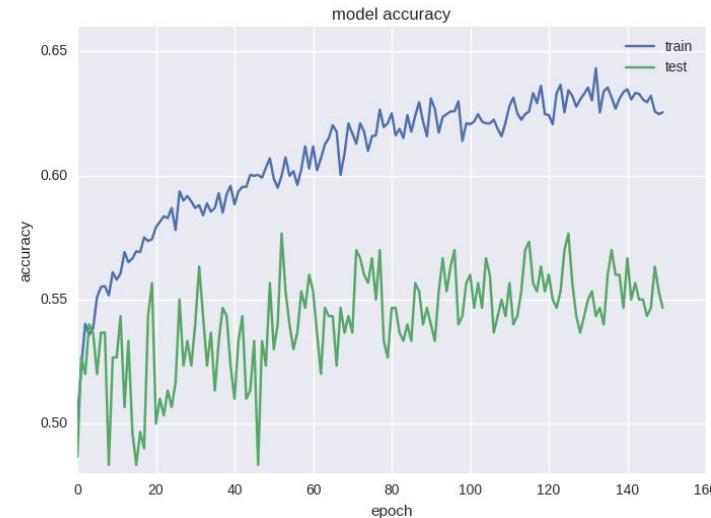
1. Часто при решении задачи классификации (движении цены) стоит решать задачу регрессии
2. Есть несколько способов предсказания на период $t+k$:
 - a. если нужно предсказать на небольшой период вперёд, по 1 модели на каждый shift
 - b. одна модель с параметром shift
 - c. одна модель на shift=1 и предсказание на основе своих предсказаний
3. Проблема: не стационарный ряд (mean, var, max, min). Тест Дики-Фуллера для проверки на стационарность

Как можно предотвратить переобучение?

```
model = Sequential()
model.add(Dense(64, input_dim=30))
model.add(BatchNormalization())
model.add(LeakyReLU())
model.add(Dense(2))
model.add(Activation('softmax'))
```



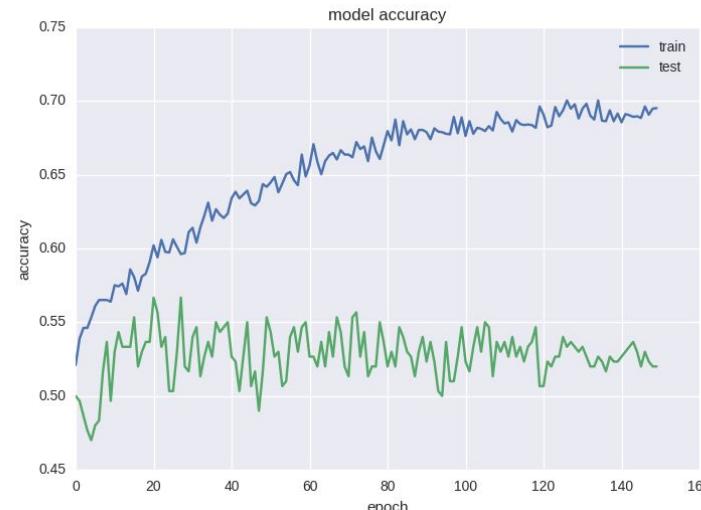
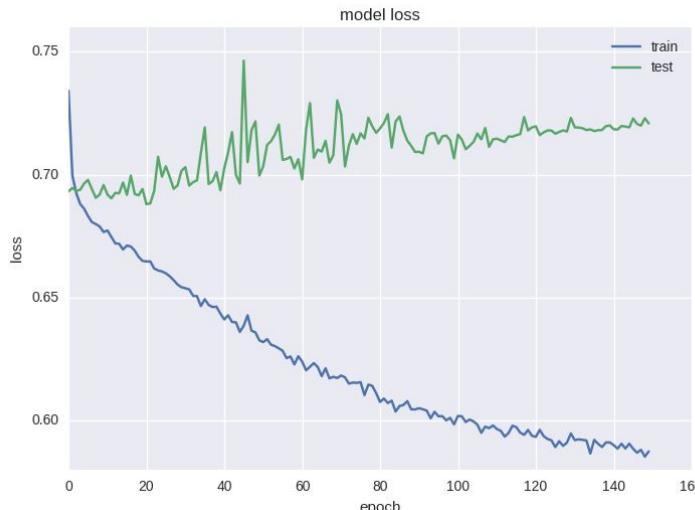
Первая попытка: обучим сеть на классификацию повышения/понижения с loss cross_entropy



Как можно предотвратить переобучение?

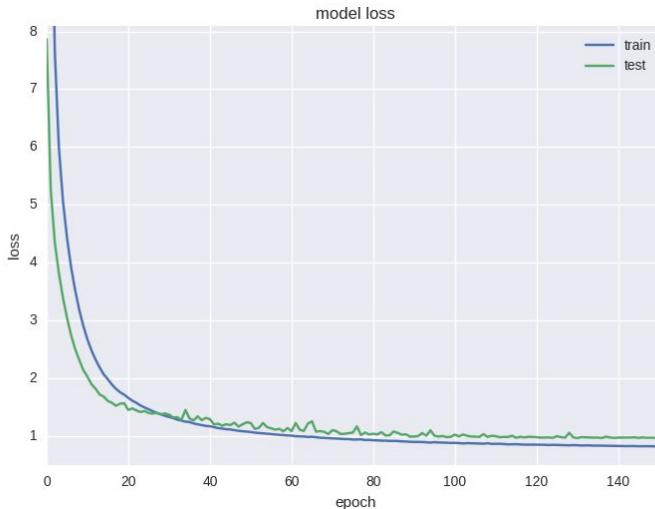
```
model = Sequential()  
model.add(Dense(64, input_dim=30))  
model.add(BatchNormalization())  
model.add(LeakyReLU())  
model.add(Dense(16))  
model.add(BatchNormalization())  
model.add(LeakyReLU())  
model.add(Dense(2)) model.add(Activation('softmax'))
```

Второй шаг: сделаем сеть глубже и посмотрим на результат



Как можно предотвратить переобучение?

```
model = Sequential()
model.add(Dense(64, input_dim=30,
activity_regularizer=regularizers.l2(0.01)))
model.add(BatchNormalization())
model.add(LeakyReLU())
model.add(Dense(16,
activity_regularizer=regularizers.l2(0.01)))
model.add(BatchNormalization())
model.add(LeakyReLU())
model.add(Dense(2))
model.add(Activation('softmax'))
```

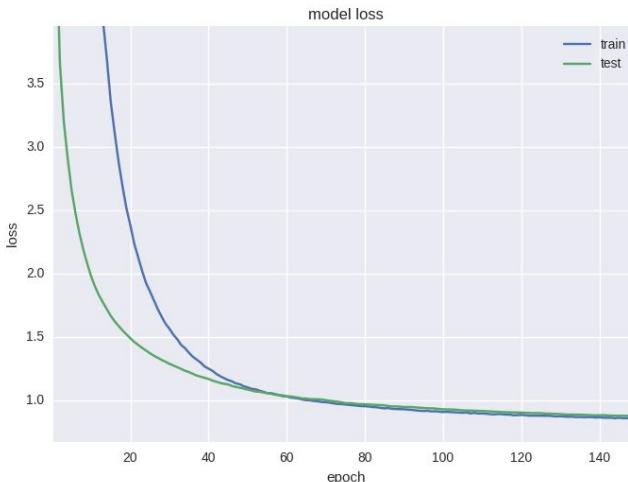


Третий шаг: добавим
регуляризацию.
Проблема: лосс падает,
качество не растёт

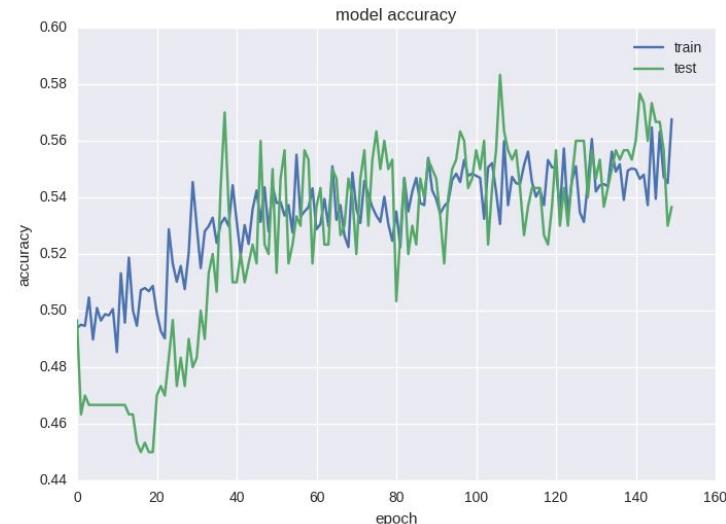


Как можно предотвратить переобучение?

```
model = Sequential()
model.add(Dense(64, input_dim=30,
activity_regularizer=regularizers.l2(0.01)))
model.add(BatchNormalization())
model.add(LeakyReLU())
model.add(Dropout(0.5))
model.add(Dense(16,
activity_regularizer=regularizers.l2(0.01)))
model.add(BatchNormalization())
model.add(LeakyReLU())
model.add(Dense(2))
model.add(Activation('softmax'))
```



Четвёртый шаг: добавим
дропаут, получили 58% accuracy



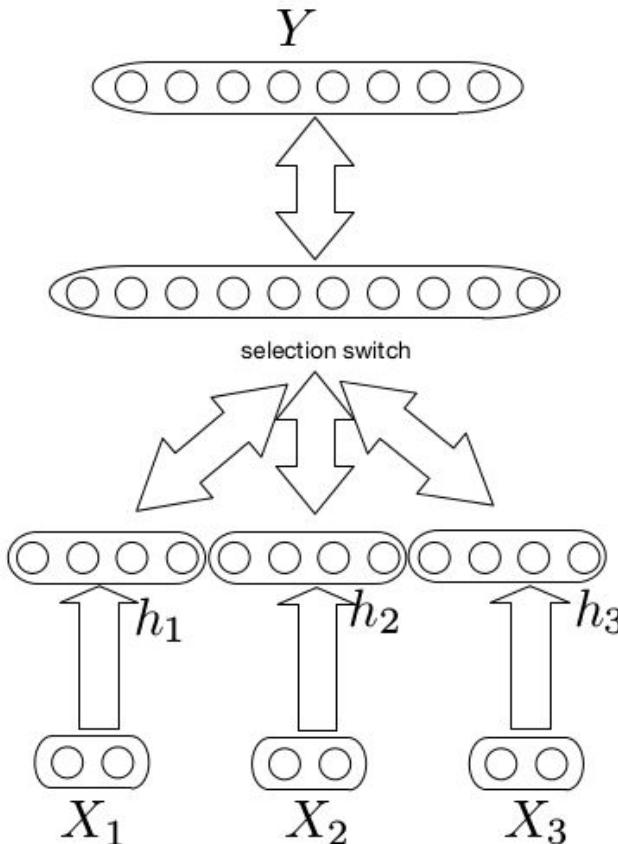
Как ещё можно обучаться и предсказывать TS?

```
main_input = Input(shape=(len(X[0]), ), name='main_input')
x = GaussianNoise(0.05)(main_input)
x = Dense(64, activation='relu')(x)
x = GaussianNoise(0.05)(x)
output = Dense(1, activation = "linear", name = "out")(x)

final_model = Model(inputs=[main_input], outputs=[output])
opt = Adam(lr=0.002)
final_model.compile(optimizer=opt, loss='mse')
```

“Novel” point here is adding small noise to the input and to the output of the single layer of our neural network. It can work very similar to L2 regularization. Proof [here](#).

Как ещё можно обучаться и предсказывать TS?

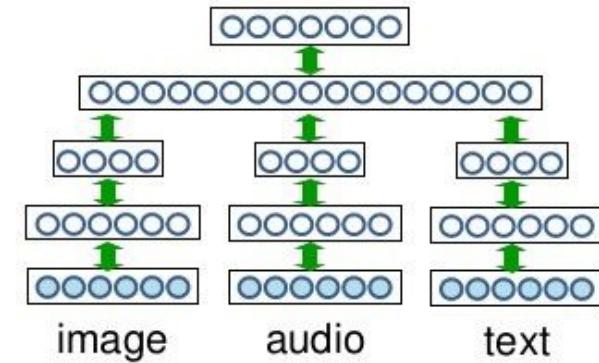


Multimodal learning model

Motivation: Multi-modal learning

Single learning algorithms that combine multiple input domains

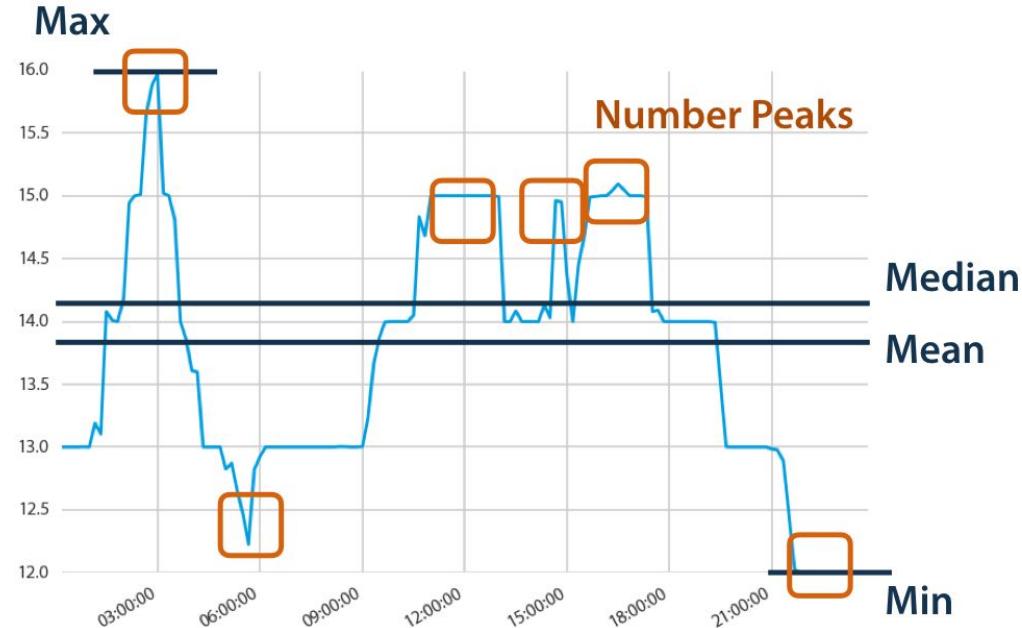
- Images
- Audio & speech
- Video
- Text
- Robotic sensors
- Time-series data
- Others



<https://becominghuman.ai/neural-networks-for-algorithmic-trading-multimodal-and-multipitask-deep-learning-5498e0098caf>

Полезные библиотеки для Time Series

Tsfresh - позволяет автоматически генерировать дополнительные признаки для временных рядов: квантили, локальные максимумы, скользящие средние, дисперсии, количество пиков и т.п.

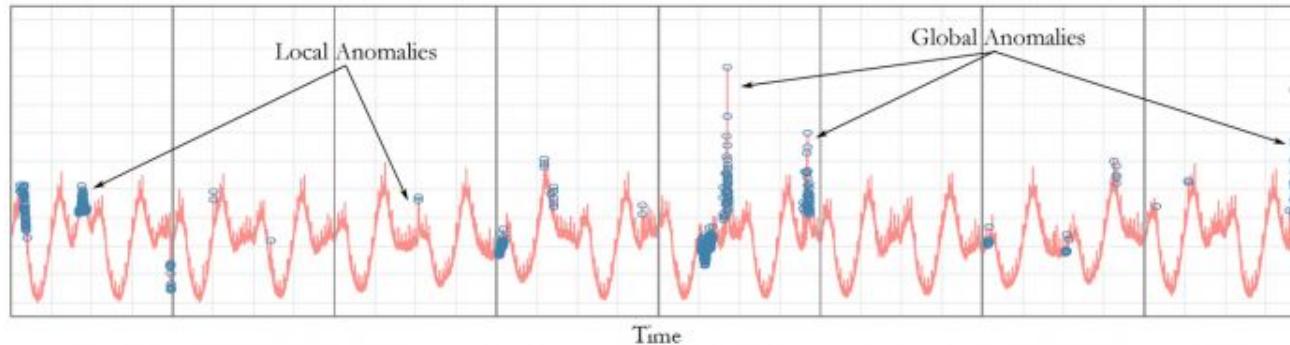
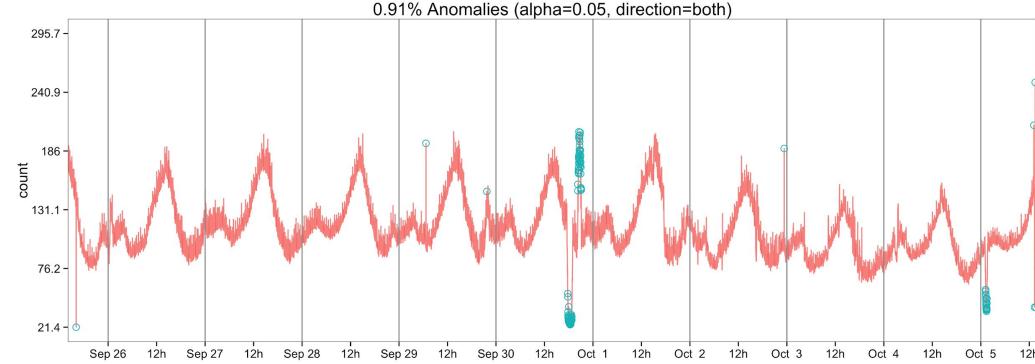


[https://github.com/blue-yonder/
tsfresh](https://github.com/blue-yonder/tsfresh)

[https://tsfresh.readthedocs.io/e
n/latest/text/list_of_features.ht
ml](https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html)

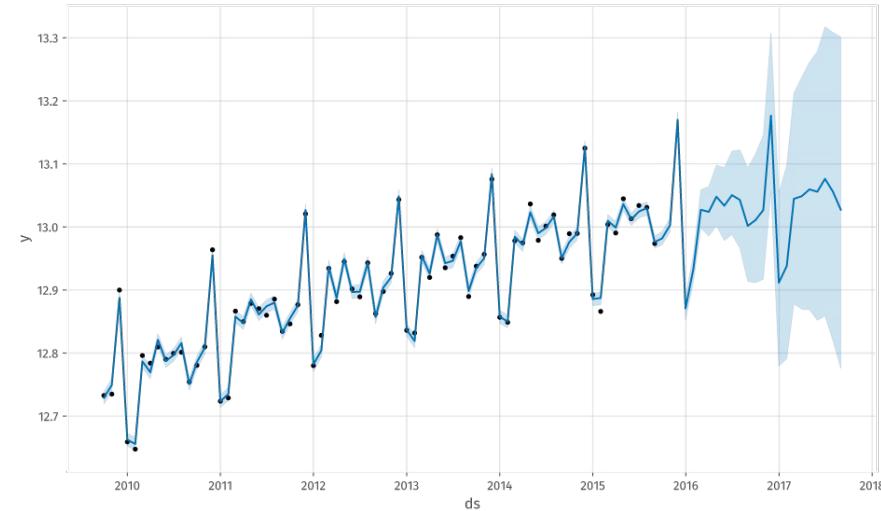
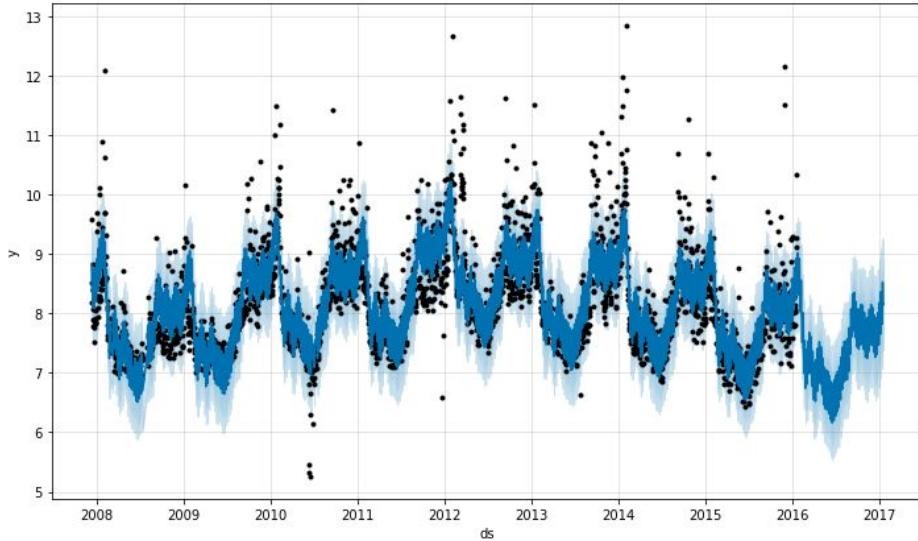
Полезные библиотеки для Time Series

Anomaly detection from Twitter - готовое решение для поиска аномалий во временных рядах, есть настройки разной чувствительности.



Полезные библиотеки для Time Series

Prophet - библиотека facebook для предсказания временных рядов. Неплохо работает из коробки.

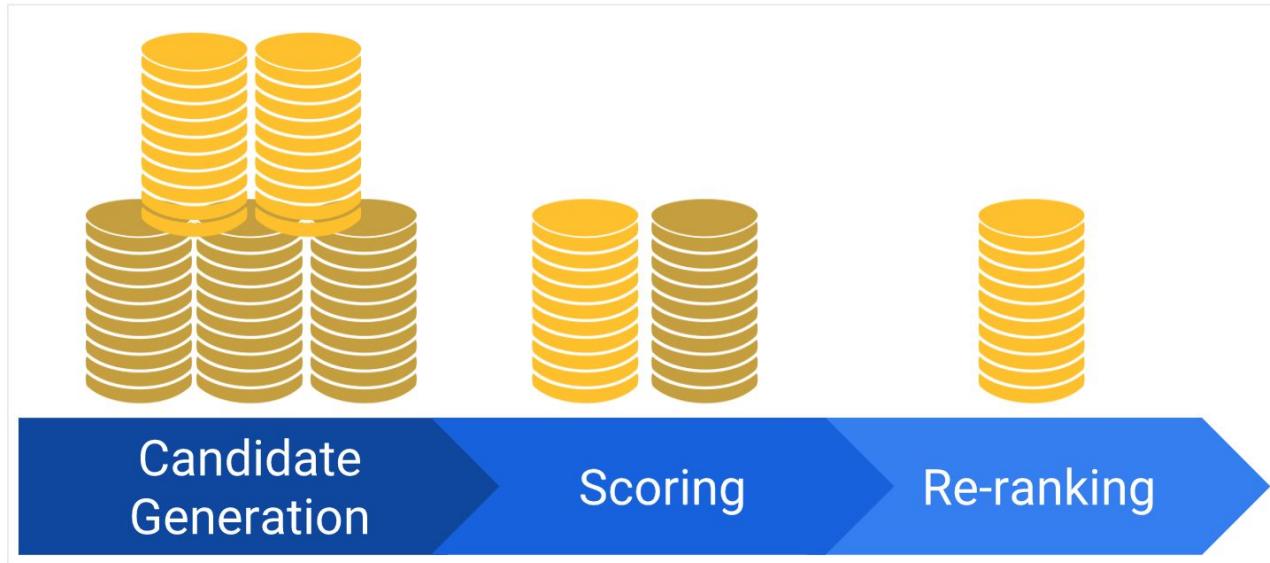


<https://facebook.github.io/prophet/>

Recommender systems

Современные рекомендательные системы состоят из
3-х основных компонент:

- candidate generation
- scoring
- re-ranking

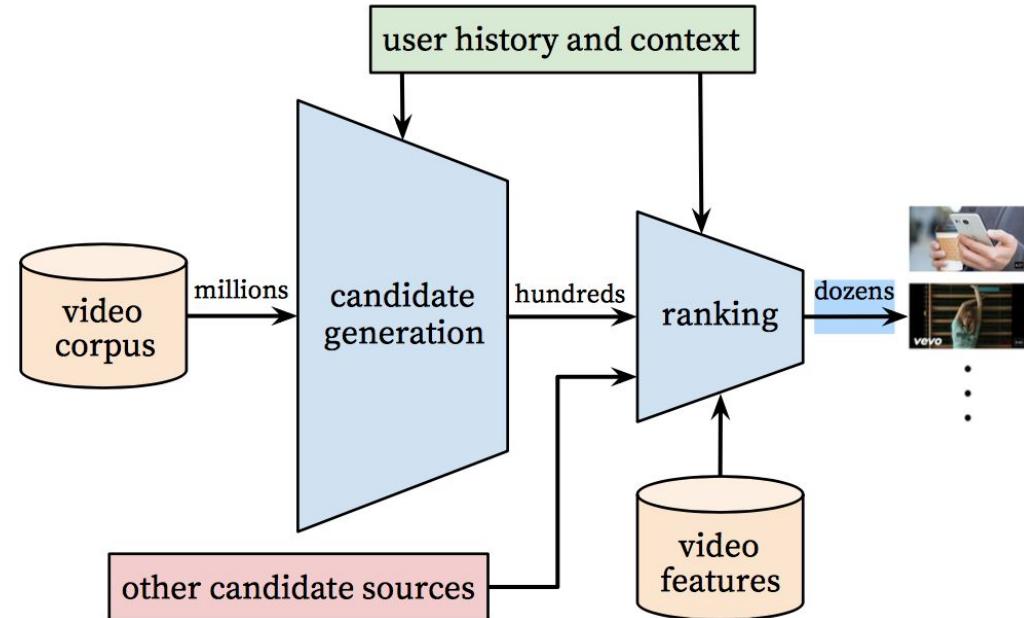


Recommender systems

Candidate generation - the system starts from a potentially huge corpus and generates a much smaller subset of candidates.

Scoring - another model scores and ranks the candidates in order to select the set of items (on the order of 10) to display to the user.

Re-ranking - must take into account additional constraints for the final ranking.



Candidate Generation

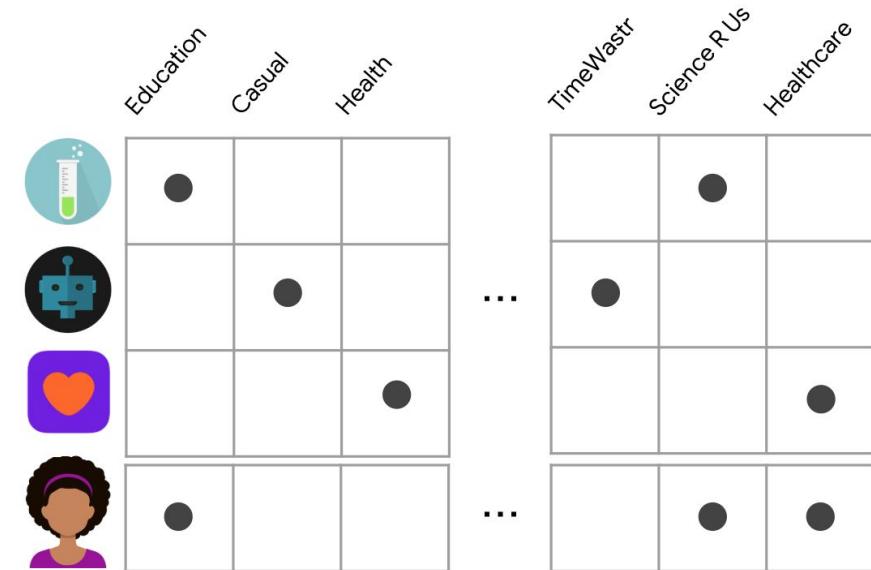
Type	Definition	Example
content-based filtering	Uses <i>similarity between items</i> to recommend items similar to what the user likes.	If user A watches two cute cat videos, then the system can recommend cute animal videos to that user.
collaborative filtering	Uses <i>similarities between queries and items simultaneously</i> to provide recommendations.	If user A is similar to user B, and user B likes video 1, then the system can recommend video 1 to user A (even if user A hasn't seen any videos similar to video 1).

Both content-based and collaborative filtering map each item and each query (or context) to an embedding vector in a common embedding space $E = \mathbb{R}^d$.

A similarity measure is a function $s : E \times E \rightarrow \mathbb{R}$ that takes a pair of embeddings and returns a scalar measuring their similarity. The embeddings can be used for candidate generation as follows: given a query embedding $q \in E$, the system looks for item embeddings $x \in E$ that are close to q , that is, embeddings with high similarity $s(q, x)$.

Candidate Generation: Content-based Filtering

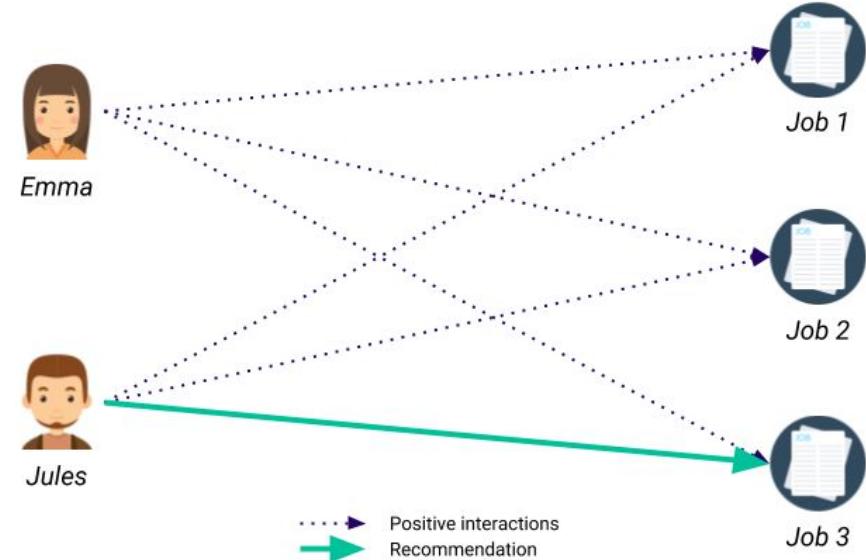
- The model doesn't need any data about other users, since the recommendations are specific to this user. This makes it easier to scale to a large number of users.
- The model can capture the specific interests of a user, and can recommend niche items that very few other users are interested in.
- Since the feature representation of the items are hand-engineered to some extent, this technique requires a lot of domain knowledge. Therefore, the model can only be as good as the hand-engineered features.
- The model can only make recommendations based on existing interests of the user. In other words, the model has limited ability to expand on the users' existing interests.



<https://developers.google.com/machine-learning/recommendation/content-based/basics>

Candidate Generation: Collaborative Filtering

- We don't need domain knowledge because the embeddings are automatically learned.
- The model can help users discover new interests.
- To some extent, the system needs only the feedback matrix to train a matrix factorization model.
- Cold start problem
- Hard to include side features for query/item



<https://www.welcometothejungle.co/en/articles/collaborative-filtering-job-recommendations>

Candidate Generation: Matrix Factorization



Observed Only MF

1			1	1	
	1				1

$$\sum_{(i, j) \in \text{obs}} (A_{ij} - U_i \cdot V_j)^2$$

Weighted MF

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

$$\sum_{(i, j) \in \text{obs}} (A_{ij} - U_i \cdot V_j)^2 + w_0 \sum_{(i, j) \notin \text{obs}} (0 - U_i \cdot V_j)^2$$

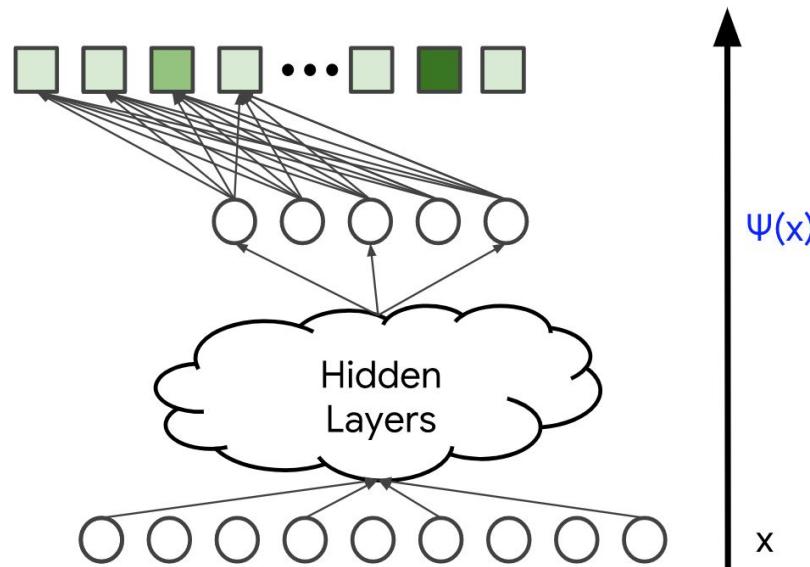
SVD

1	0	1	1	0
0	1	0	0	1
1	1	1	0	0
0	0	0	1	1

$$= \sum_{(i, j)} |A - UV^T|_F^2 = \sum_{(i, j)} (A_{ij} - U_i \cdot V_j)^2$$

Candidate Generation: Deep Learning

Deep neural network (DNN) models can address these limitations of matrix factorization. DNNs can easily incorporate query features and item features (due to the flexibility of the input layer of the network), which can help capture the specific interests of a user and improve the relevance of recommendations.

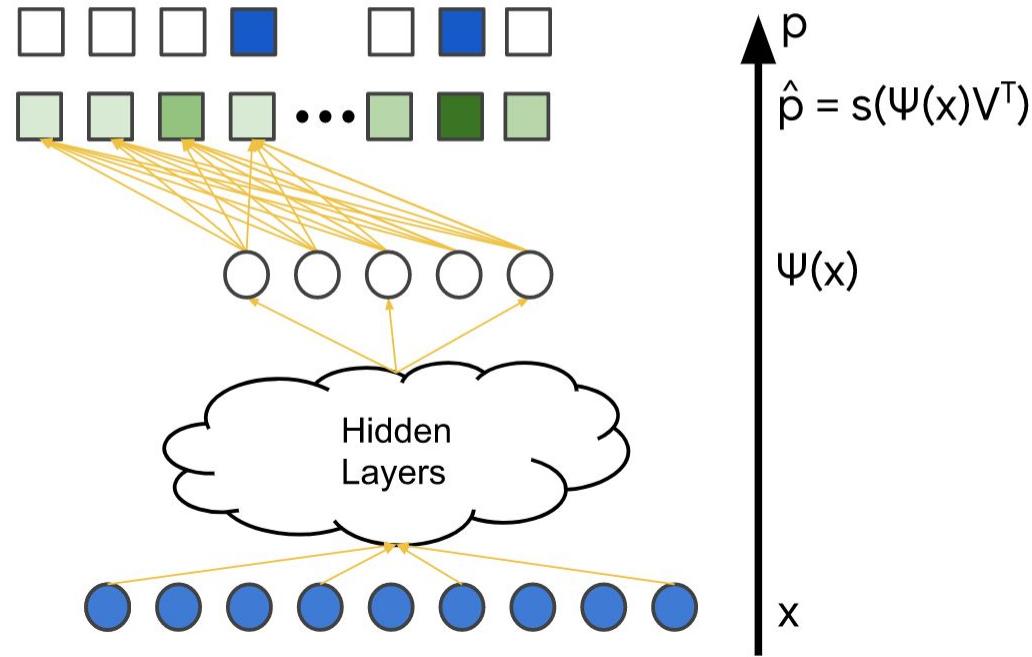


One possible DNN model is **softmax**, which treats the problem as a multiclass prediction problem in which:

- The input is the user query.
- The output is a probability vector with size equal to the number of items in the corpus, representing the probability to interact with each item; for example, the probability to click on or watch a YouTube video.

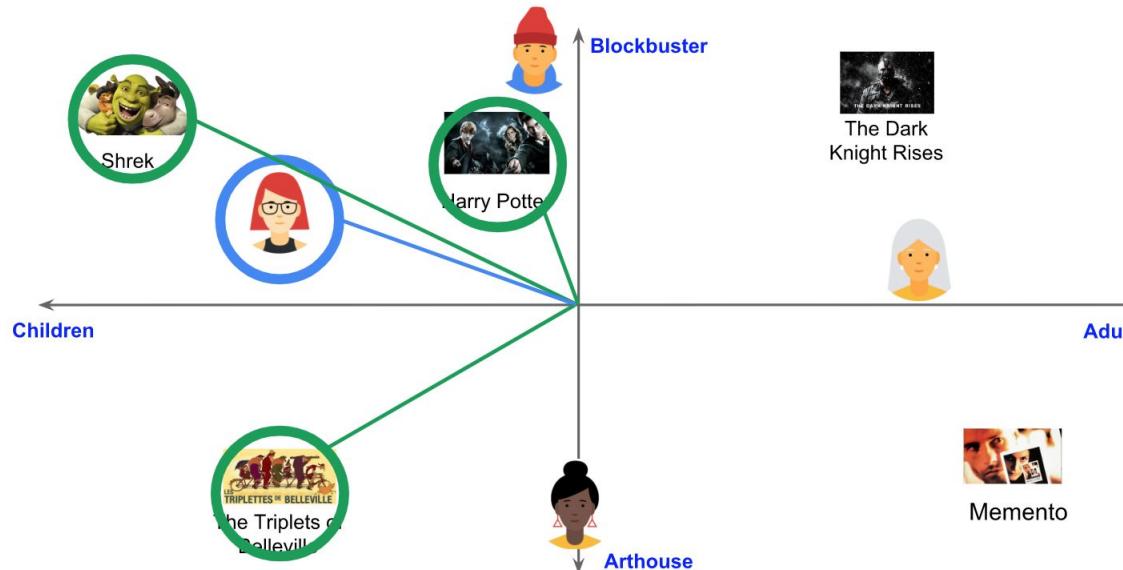
Candidate Generation: Deep Learning

The softmax training data consists of the query features x and a vector of items the user interacted with (represented as a probability distribution p). These are marked in blue in the following figure. The variables of the model are the weights in the different layers. These are marked as orange in the following figure. The model is typically trained using any variant of stochastic gradient descent.



Scoring

Once you have the query embedding q , search for item embeddings V_j that are close to q in the embedding space. This is a nearest neighbor problem. For example, you can return the top k items according to the similarity score $s(q, V_j)$.



To compute the nearest neighbors in the embedding space, the system can exhaustively score every potential candidate. Exhaustive scoring can be expensive for very large corpora, but you can use either of the following strategies to make it more efficient:

- If the query embedding is known statically, the system can perform exhaustive scoring offline, precomputing and storing a list of the top candidates for each query. This is a common practice for related-item recommendation.
- Use approximate nearest neighbors.

Re-ranking

You can implement re-ranking on a video recommender by doing the following:

1. Training a separate model that detects whether a video is click-bait.
2. Running this model on the candidate list.
3. Removing the videos that the model classifies as click-bait.

The system re-ranks videos by modifying the score as a function of:

- video age (perhaps to promote fresher content)
- video length
- Freshness
- Diversity



Cute Owl video
Owler
Recommended for you



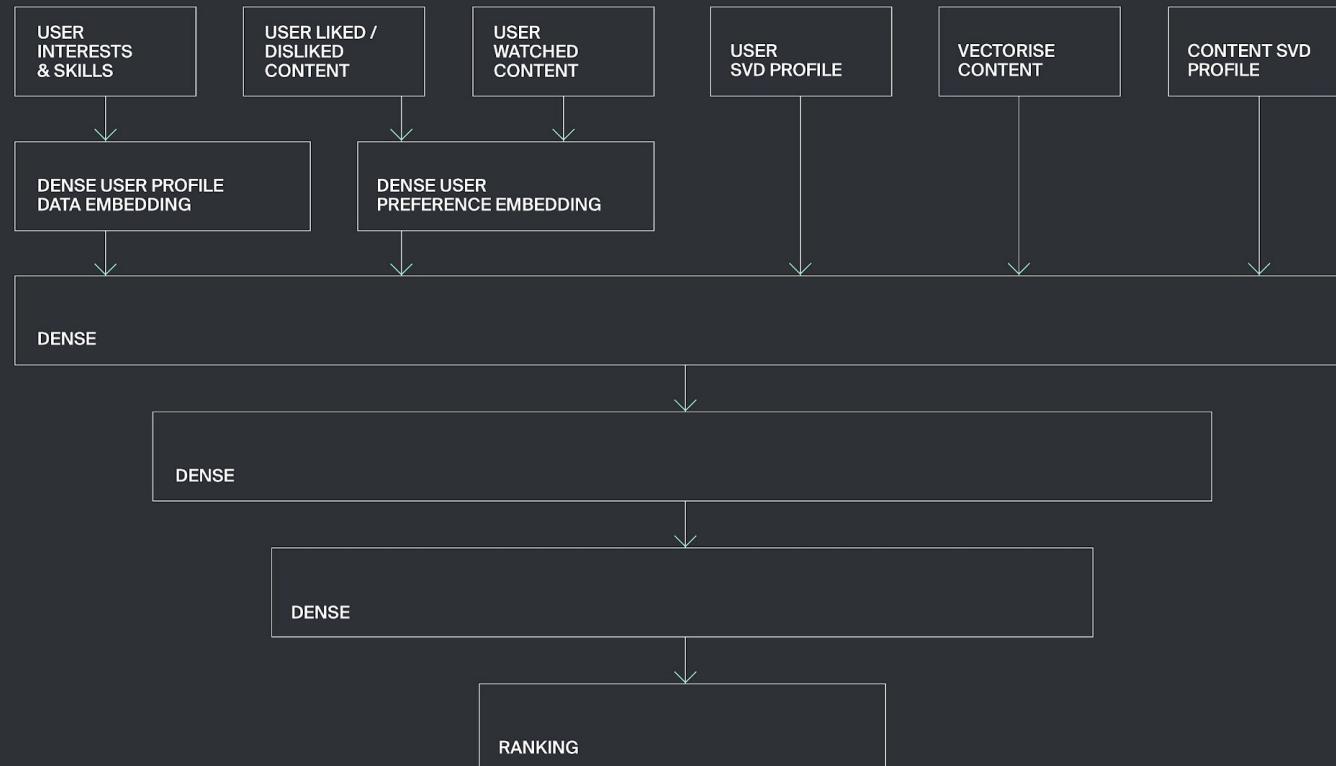
Cute Owlet video
Bird Watcher
Recommended for you

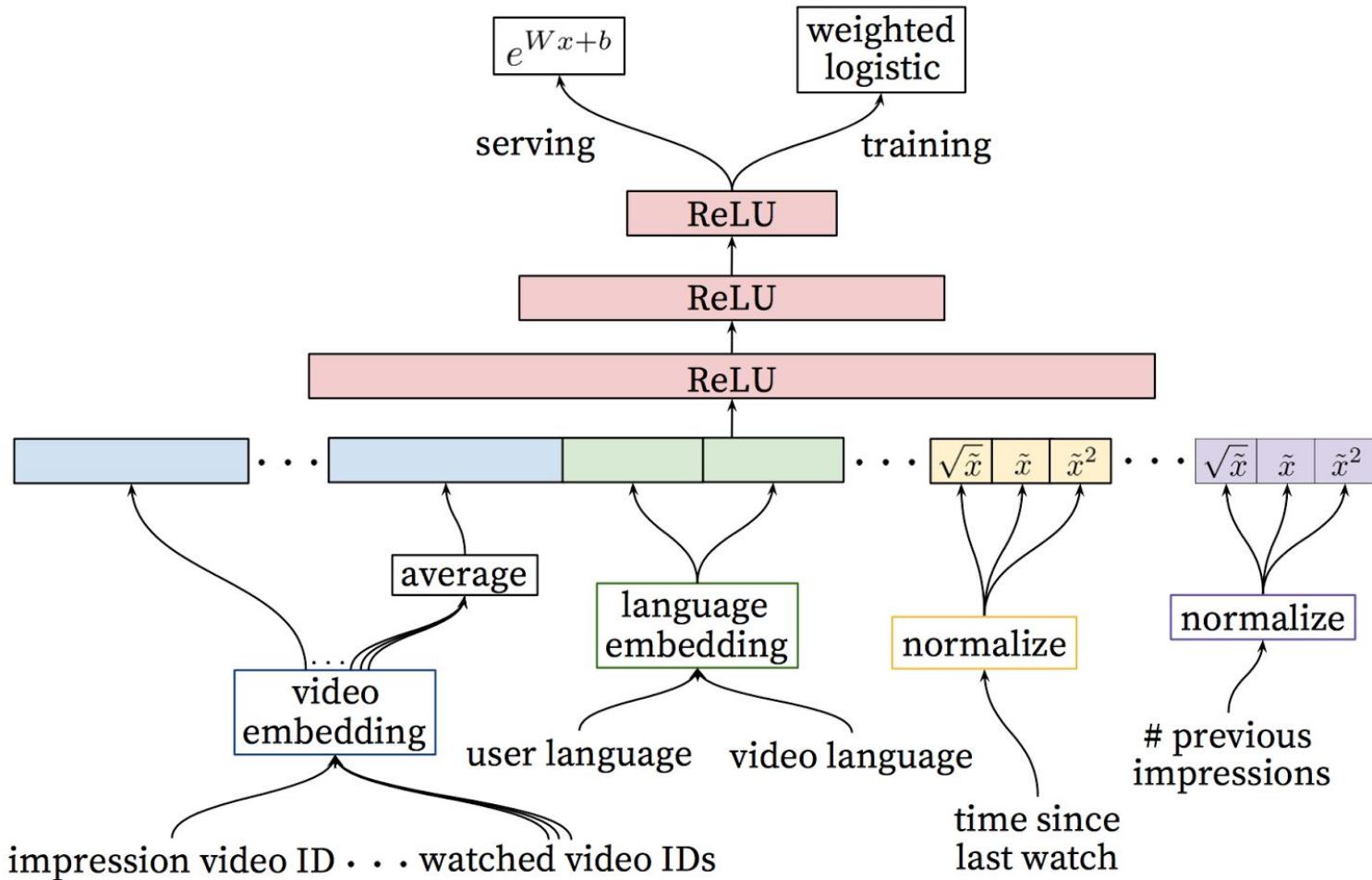


Another Owl video
Birdz
Recommended for you



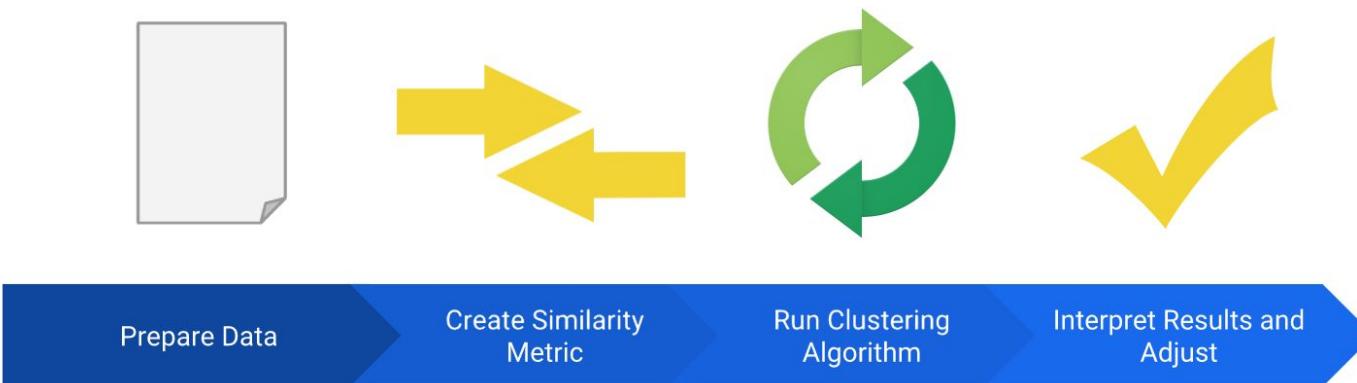
Owl video
Granny G
Recommended for you





Clustering

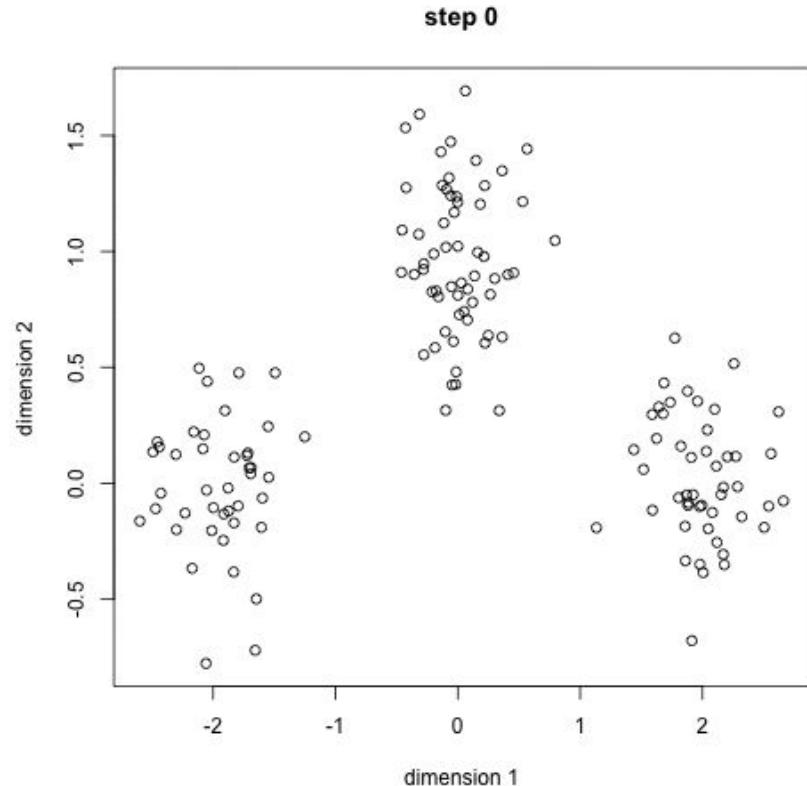
1. As with any ML problem, you must normalize, scale, and transform feature data. While clustering however, you must additionally ensure that the prepared data lets you accurately calculate the similarity between examples
2. Before a clustering algorithm can group data, it needs to know how similar pairs of examples are. You quantify the similarity between examples by creating a similarity metric. Creating a similarity metric requires you to carefully understand your data and how to derive similarity from your features
3. A clustering algorithm uses the similarity metric to cluster data



Clustering: k-means

- 1) Инициализируем центры случайными объектами
- 2) Относим каждый объект к ближайшему из центров и присваиваем ему соответствующий класс
- 3) Двигаем центры в центры масс получившихся кластеров
- 4) Goto 2) пока сдвиг на 3) $\geq \text{eps}$

Одна итерация за $O(nk)$



<http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html>

Clustering: EM

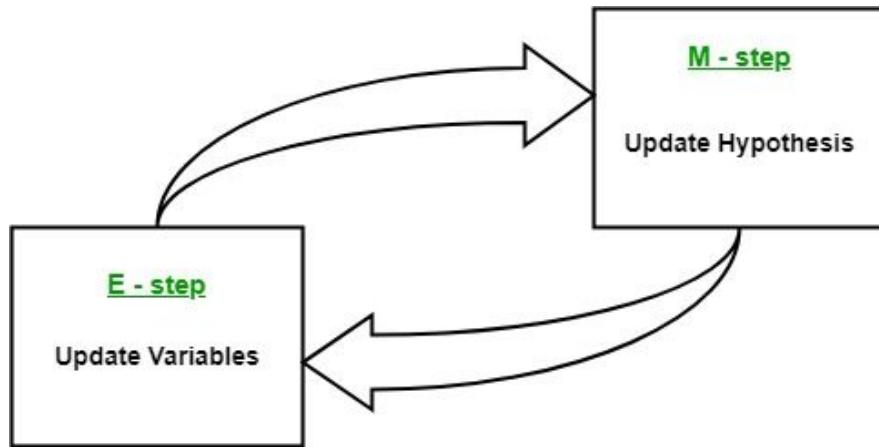
- Observations $x_1 \dots x_n$
 - K=2 Gaussians with unknown μ, σ^2
- Chicken and egg problem
 - need (μ_a, σ_a^2) and (μ_b, σ_b^2) to guess source of points
 - need to know source to estimate (μ_a, σ_a^2) and (μ_b, σ_b^2)
- EM algorithm
 - start with two randomly placed Gaussians $(\mu_a, \sigma_a^2), (\mu_b, \sigma_b^2)$

E-step: – for each point: $P(b|x_i) =$ does it look like it came from b?

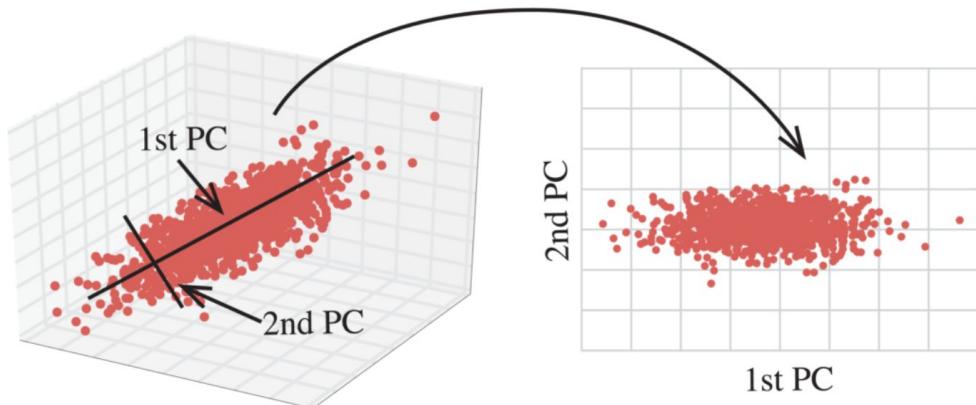
M-step: – adjust (μ_a, σ_a^2) and (μ_b, σ_b^2) to fit points assigned to them

- iterate until convergence

Clustering: EM



Dimensionality Reduction: PCA

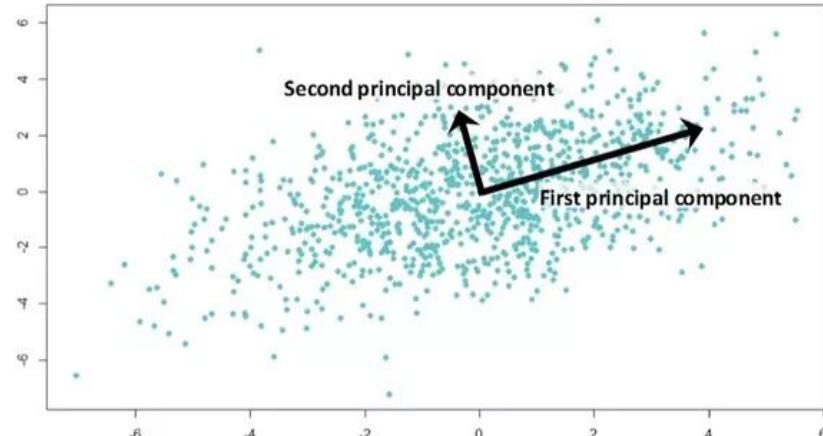


Один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.



Dimensionality Reduction: PCA

- 1) Вычисляем матрицу ковариаций признаков
- 2) Находим собственные вектора матрицы ковариаций
- 3) Первые k векторов соответствующих к максимальным собственным значениям - компоненты нашего разложения



<https://medium.com/@sadatnazrul/the-dos-and-donts-of-principal-component-analysis-7c2e9dc8cc48>

Плюсы: возможность регуляции получаемой размерности (добавлении компонент по одной в зависимости от объяснённой дисперсии); скорость алгоритма; интерпретация

Минусы: линейность, предположение об ортогональности

Dimensionality Reduction: t-SNE

У нас есть набор данных с точками, описываемыми многомерной переменной с размерностью пространства существенно больше трех. Необходимо получить новую переменную, существующую в двумерном или трехмерном пространстве, которая бы в максимальной степени сохраняла структуру и закономерности в исходных данных.

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Эта формула показывает, насколько точка X_j близка к точке X_i при гауссовом распределении вокруг X_i с заданным отклонением σ . Сигма будет различной для каждой точки. Она выбирается так, чтобы точки в областях с большей плотностью имели меньшую σ .

Для двумерных или трехмерных «коллег» пары X_i и X_j , назовем их для ясности Y_i и Y_j , не представляя труда оценить условную вероятность, используя ту же формулу 1.

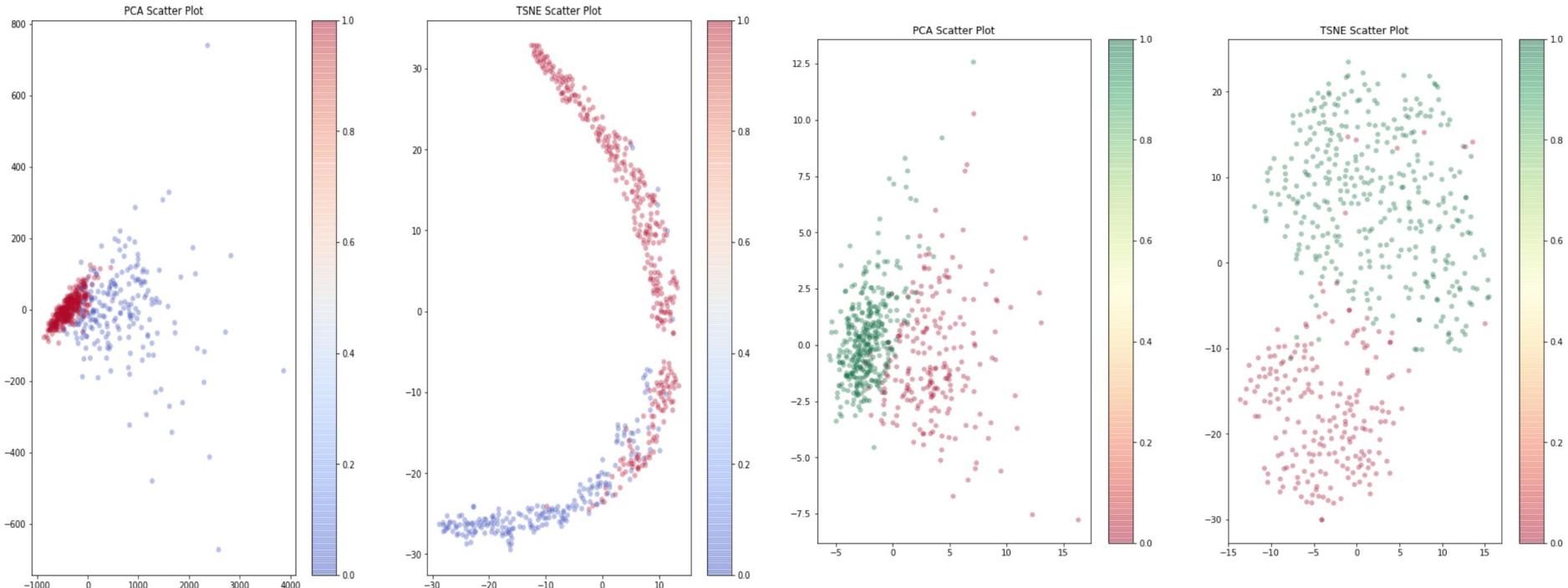
Стандартное отклонение предлагается установить в $1/\sqrt{2}$:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

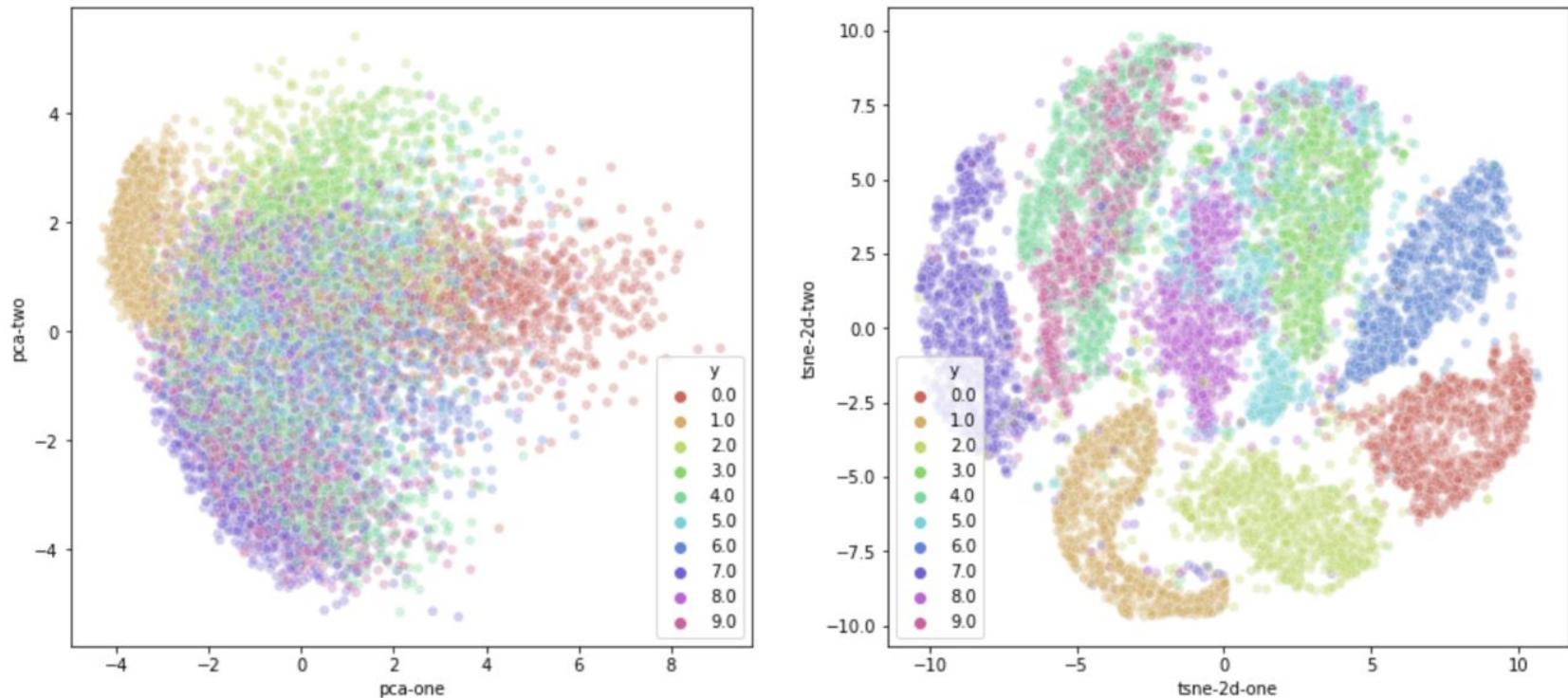
$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n},$$

$$Cost = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \quad Cost = KL(P \| Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

Dimensionality Reduction: PCA vs t-SNE



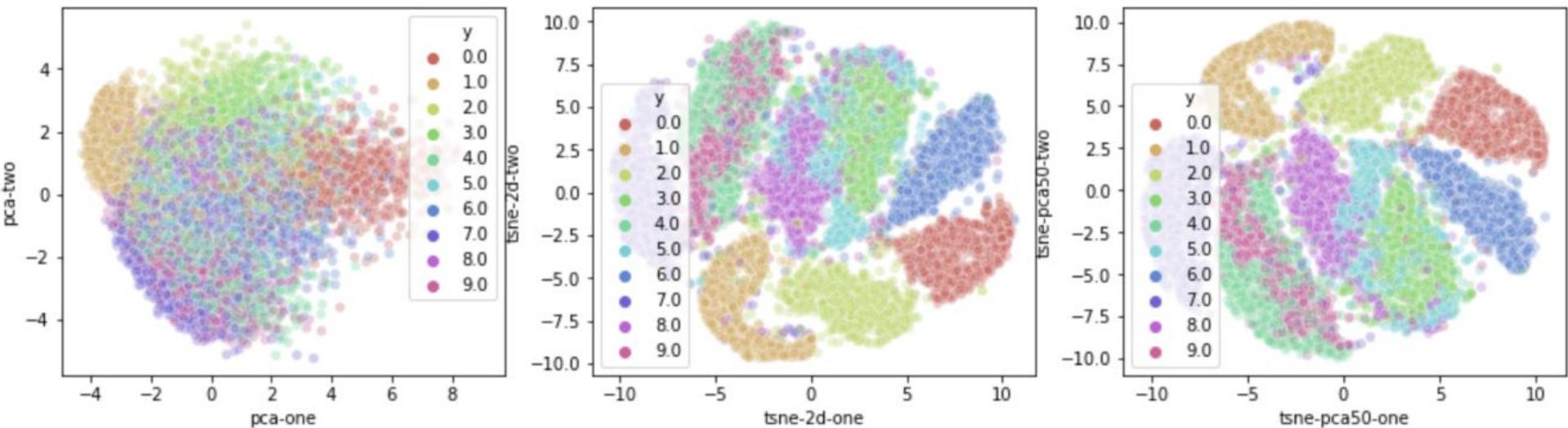
Dimensionality Reduction: PCA vs t-SNE



PCA (left) vs T-SNE (right)

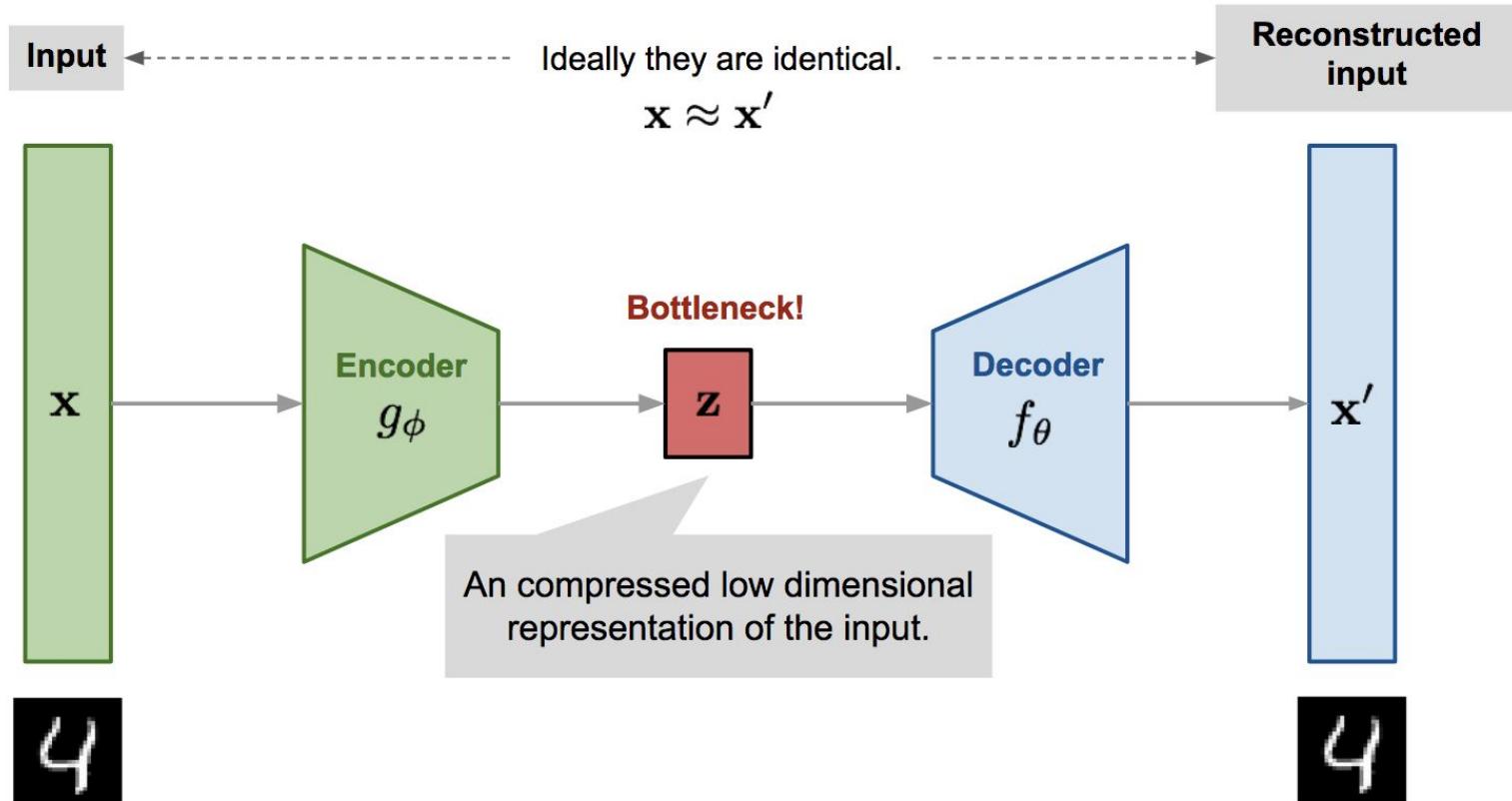
<https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b>

Dimensionality Reduction: PCA vs t-SNE



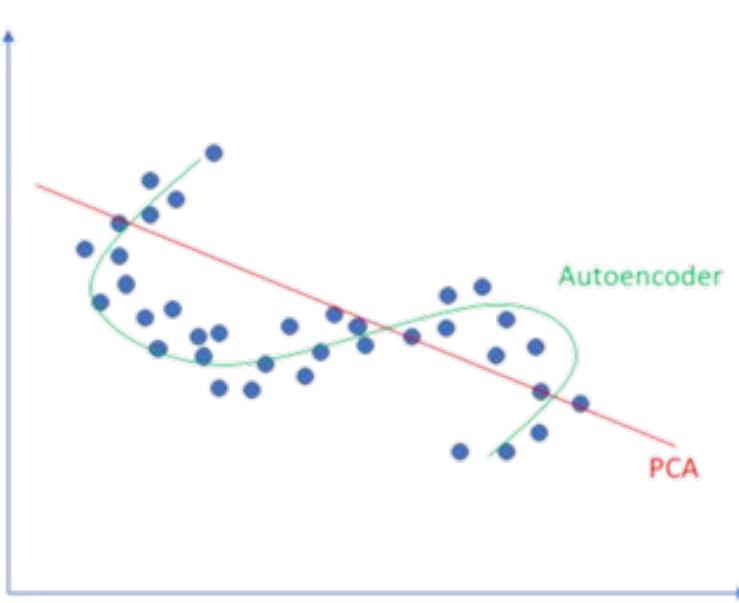
PCA (left) vs T-SNE (middle) vs T-SNE on PCA50 (right)

Dimensionality Reduction: Autoencoder



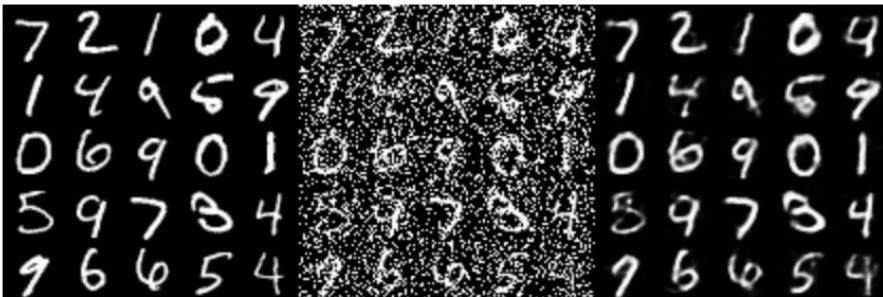
Dimensionality Reduction: Autoencoder

Linear vs nonlinear dimensionality reduction



- **non-linear transformations** with a **non-linear activation function** and multiple layers
- **convolutional layers** to learn which is better for video, image and series data
- It is more efficient to learn several layers with an autoencoder rather than learn one huge transformation with PCA
- **pre-trained layers** from another model to apply transfer learning to enhance the encoder/decoder

Autoencoder



The reconstructed image is the same as our input but with reduced dimensions. It helps in providing the similar image with a reduced pixel value.

Autoencoder

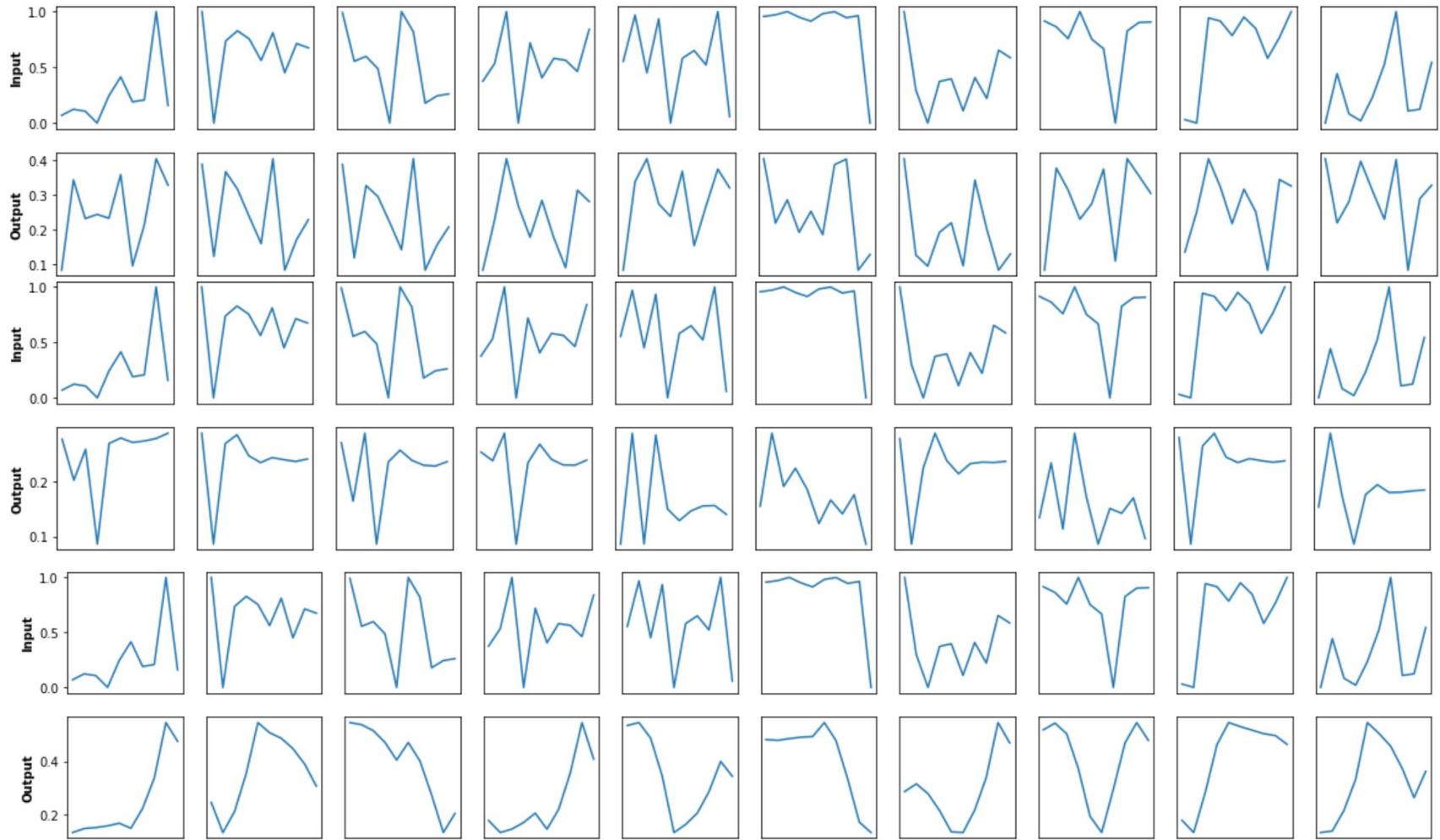
It is also used for removing watermarks from images or to remove any object while filming a video or a movie.



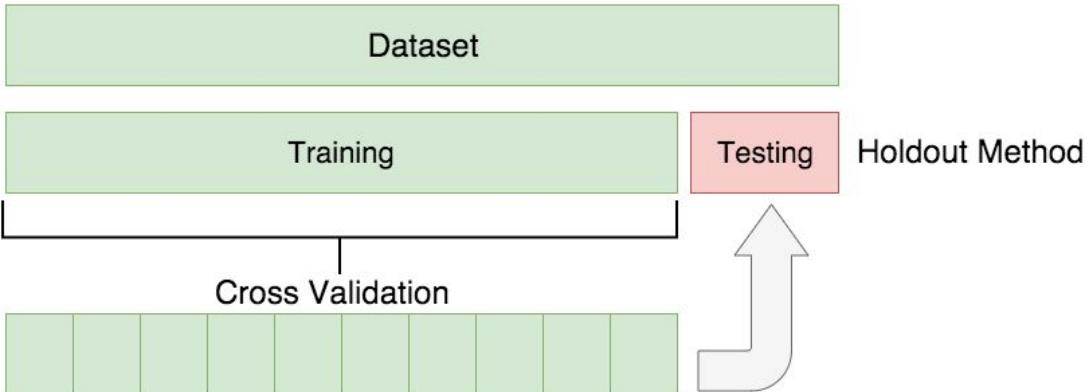
Autoencoders are used for converting any black and white picture into a colored image. Depending on what is in the picture, it is possible to tell what the color should be.



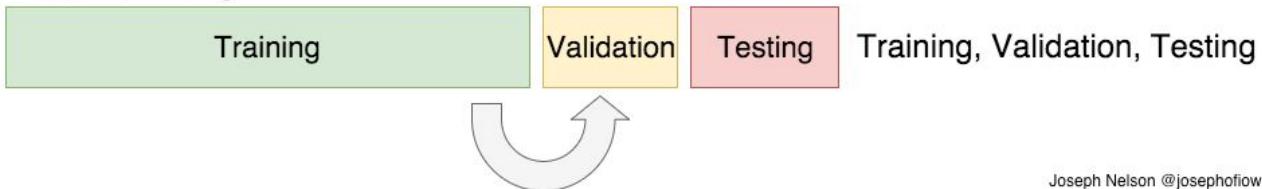
Autoencoder for Time Series?!



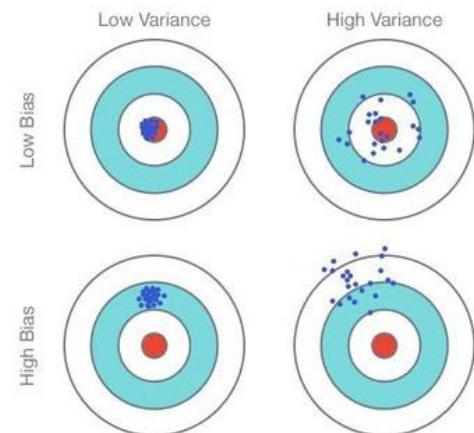
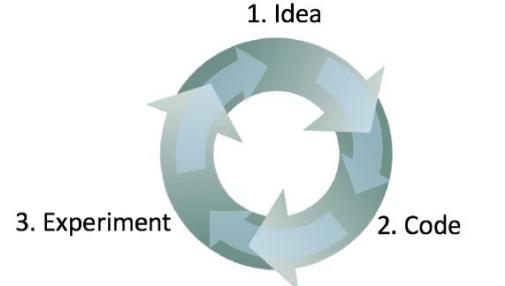
Summary - Lecture 1



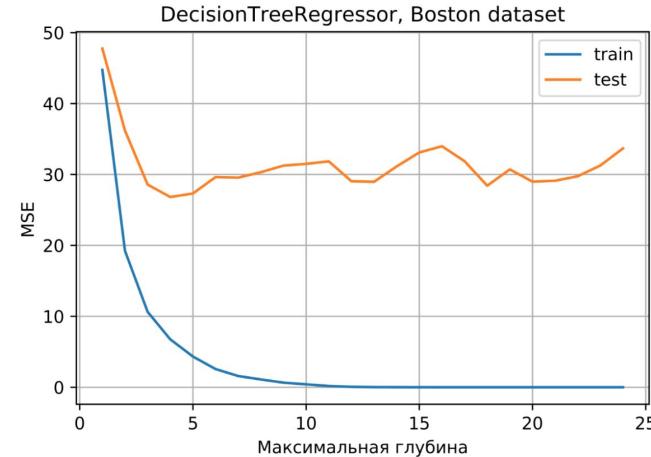
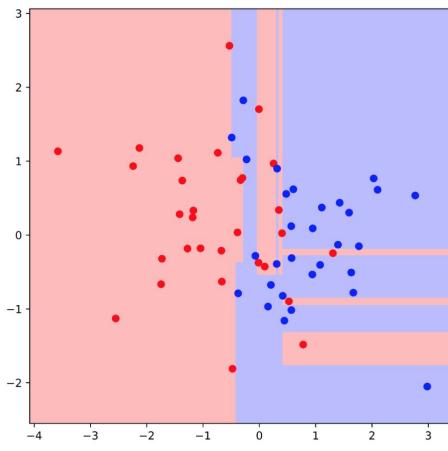
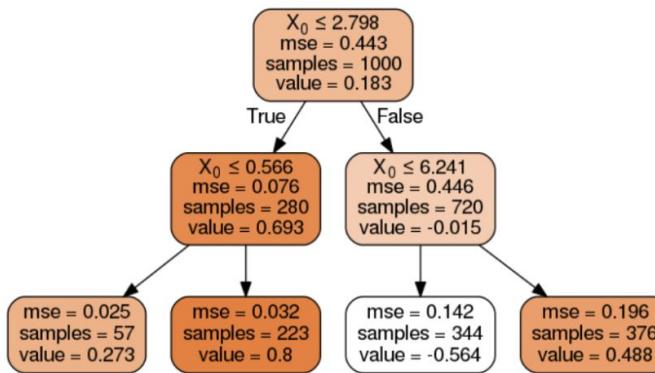
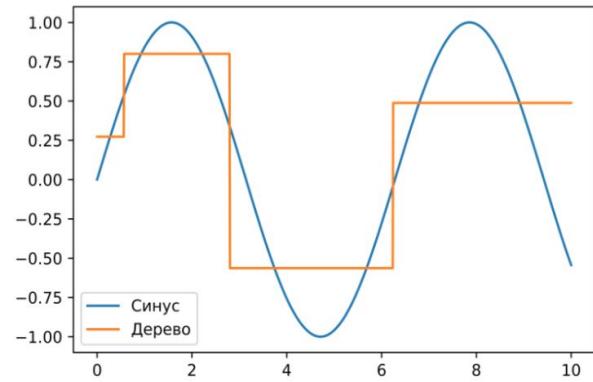
Data Permitting:



Joseph Nelson @josephofiowa



Summary - Lecture 2

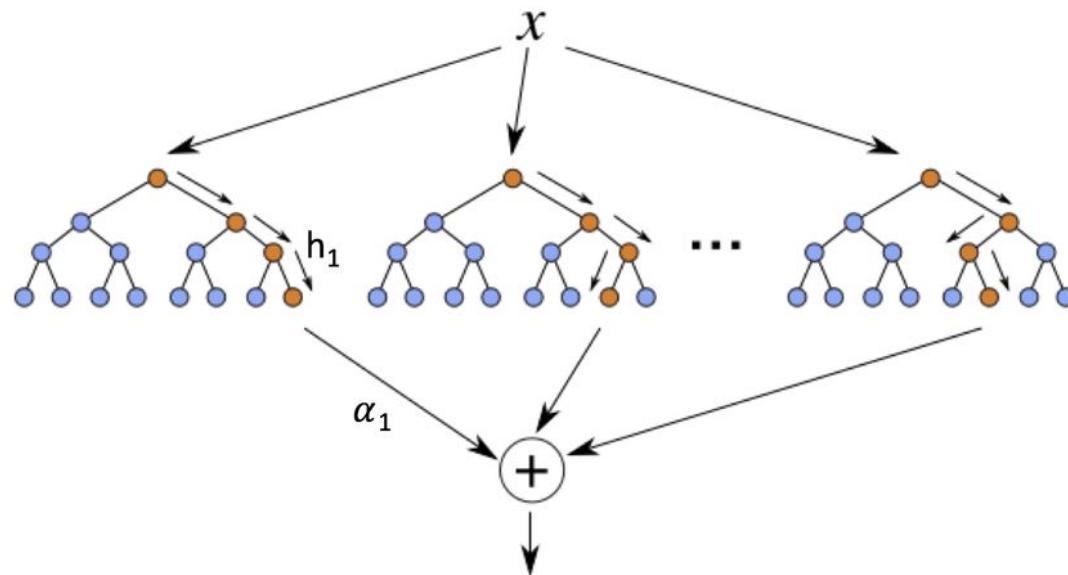


Summary - Lecture 2

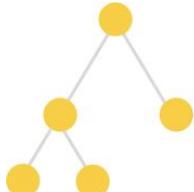
MSE: $H(Y) = \frac{1}{|Y|} \sum_{i=1}^{|Y|} (y_i - \bar{y})^2$ ■ Энтропия: $H(X) = - \sum_{i=1}^L P_i \log P_i$

$$\bar{y} = \frac{1}{|Y|} \sum_{i=1}^{|Y|} y_i$$

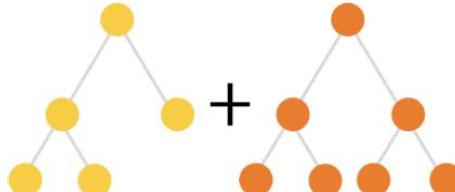
■ Джини: $H(X) = \sum_{i=1}^L P_i(1 - P_i)$



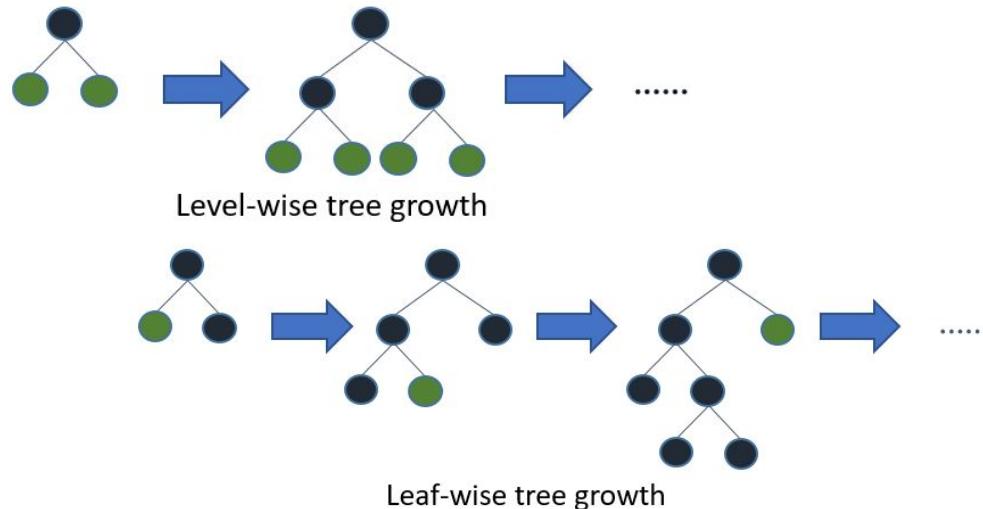
Summary - Lecture 2



First Tree

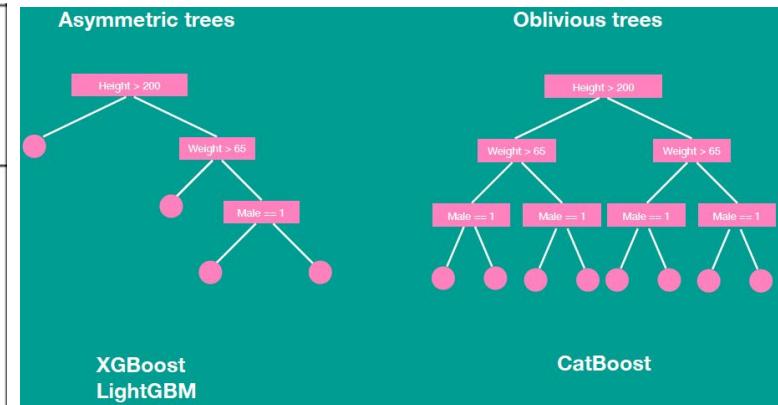


Second Tree

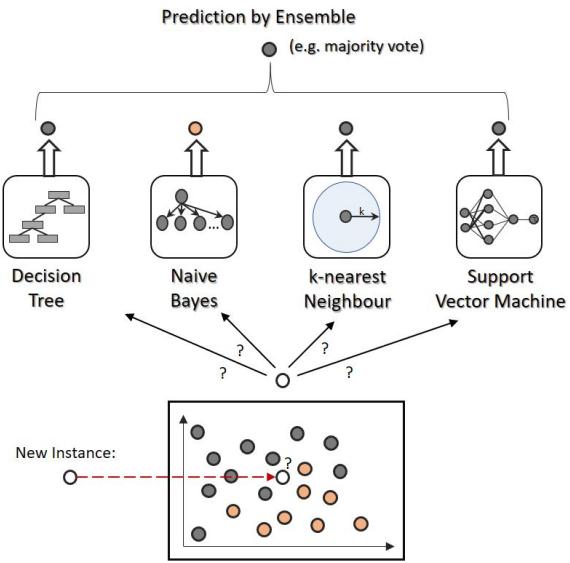


Leaf-wise tree growth

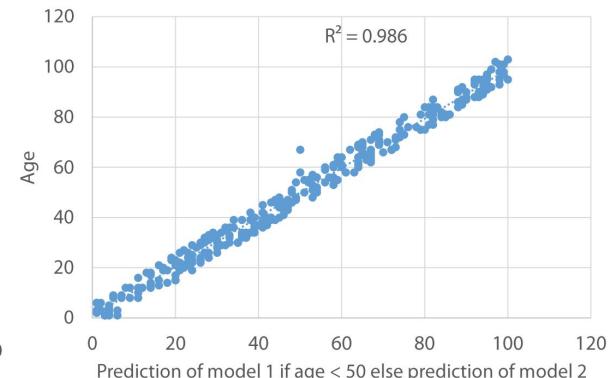
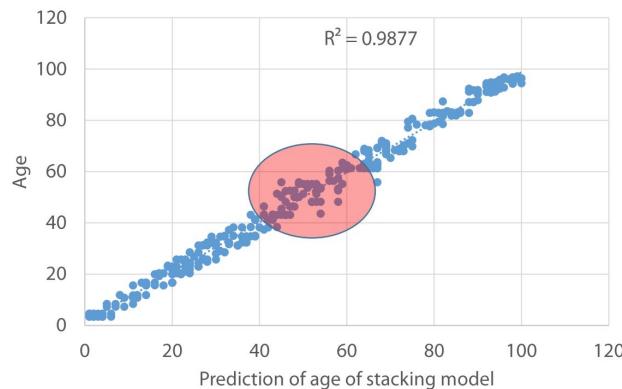
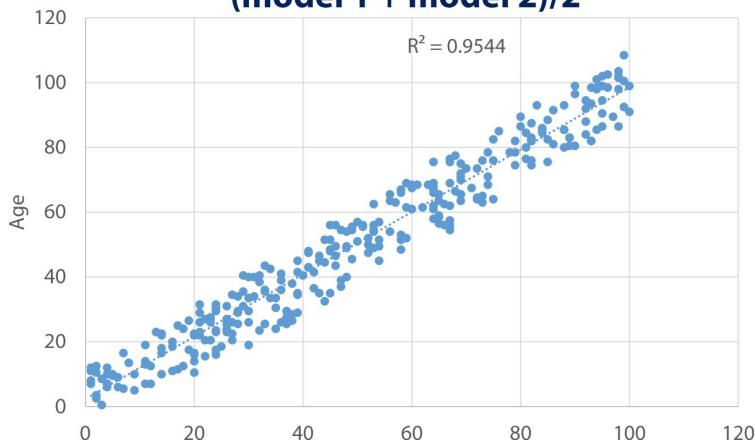
Random Forest	<ul style="list-style-type: none"> N_estimators Max_depth Min_samples_split Min_samples_leaf Max features 	<ul style="list-style-type: none"> 120, 300, 500, 800, 1200 5, 8, 15, 25, 30, None 1, 2, 5, 10, 15, 100 1, 2, 5, 10 Log2, sqrt, None
Xgboost	<ul style="list-style-type: none"> Eta Gamma Max_depth Min_child_weight Subsample Colsample_bytree Lambda alpha 	<ul style="list-style-type: none"> 0.01, 0.015, 0.025, 0.05, 0.1 0.05-0.1, 0.3, 0.5, 0.7, 0.9, 1.0 3, 5, 7, 9, 12, 15, 17, 25 1, 3, 5, 7 0.6, 0.7, 0.8, 0.9, 1.0 0.6, 0.7, 0.8, 0.9, 1.0 0.01-0.1, 1.0 , RS* 0, 0.1, 0.5, 1.0 RS*



Summary - Lecture 3



(model 1 + model 2)/2



Summary - Lecture 3

A				
x0	x1	x2	xn	y
0.17	0.25	0.93	0.79	1
0.35	0.61	0.93	0.57	0
0.44	0.59	0.56	0.46	0
0.37	0.43	0.74	0.28	1
0.96	0.07	0.57	0.01	1

B				
x0	x1	x2	xn	y
0.89	0.72	0.50	0.66	0
0.58	0.71	0.92	0.27	1
0.10	0.35	0.27	0.37	0
0.47	0.68	0.30	0.98	0
0.39	0.53	0.59	0.18	1

C				
x0	x1	x2	xn	y
0.29	0.77	0.05	0.09	?
0.38	0.66	0.42	0.91	?
0.72	0.66	0.92	0.11	?
0.70	0.37	0.91	0.17	?
0.59	0.98	0.93	0.65	?

Train algorithm **0** on A and make predictions for B and C and save to **B1, C1**

Train algorithm **1** on A and make predictions for B and C and save to **B1, C1**

Train algorithm **2** on A and make predictions for B and C and save to **B1, C1**

B1			
pred0	pred1	pred2	y
0.24	0.72	0.70	0
0.95	0.25	0.22	1
0.64	0.80	0.96	0
0.89	0.58	0.52	0
0.11	0.20	0.93	1

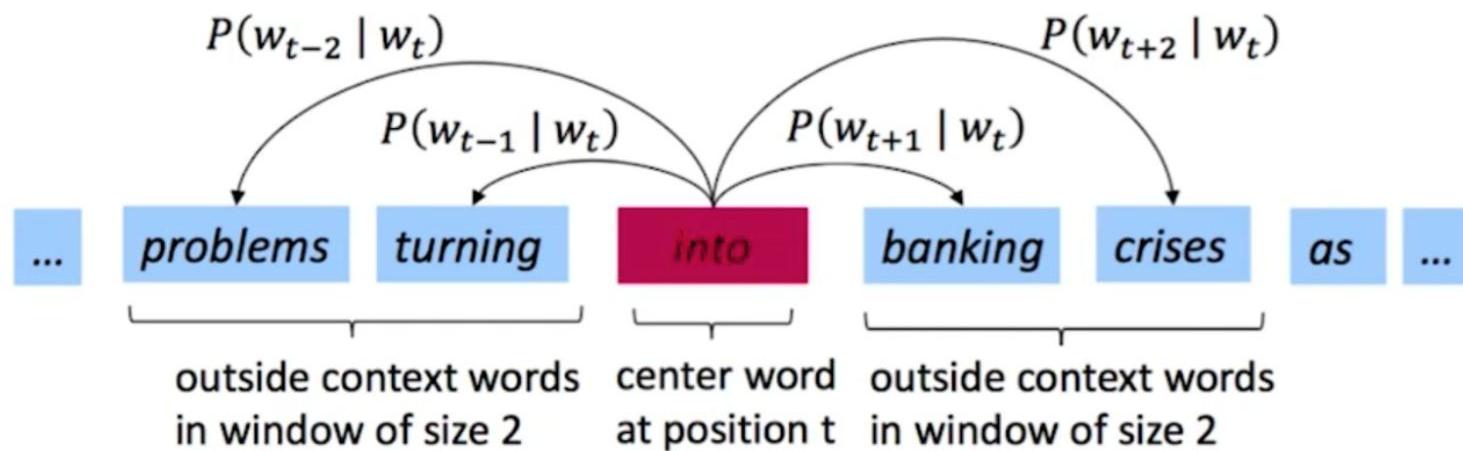
C1				
pred0	pred1	pred2	y	Preds3
0.50	0.50	0.39	?	0.45
0.62	0.59	0.46	?	0.23
0.22	0.31	0.54	?	0.99
0.90	0.47	0.09	?	0.34
0.20	0.09	0.61	?	0.05

Train algorithm **3** on B1 and make predictions for C1

Summary - Lecture 3

Word2Vec

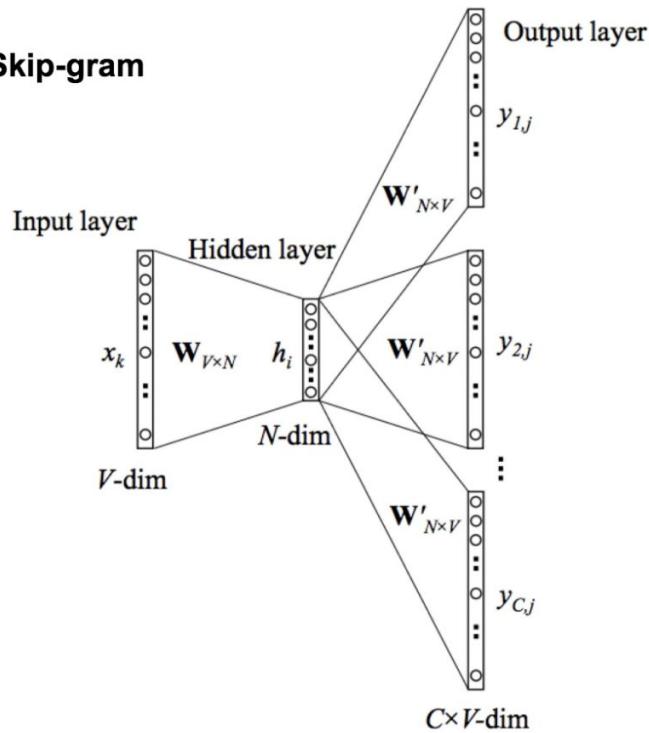
Going through text corpus by sliding windows



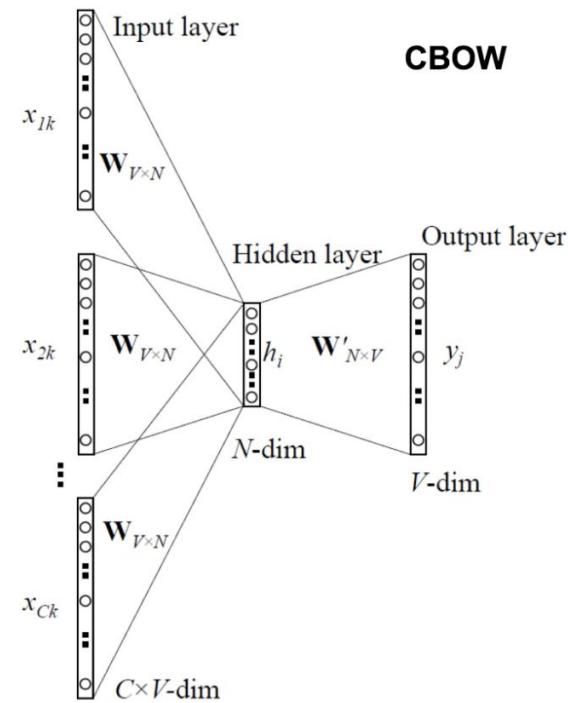
Summary - Lecture 3

Word2Vec

Skip-gram

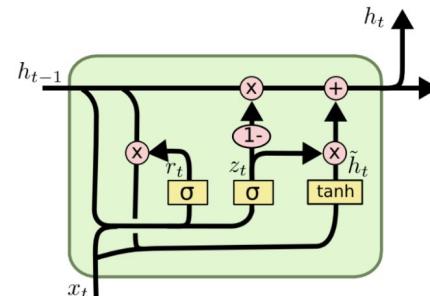
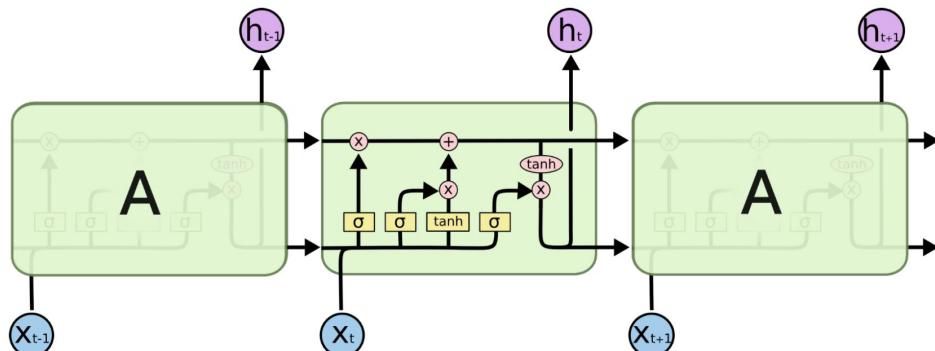
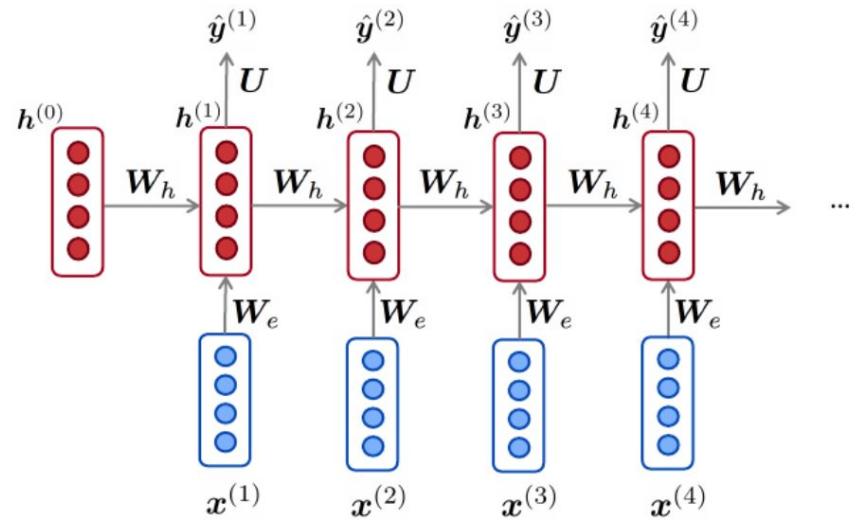
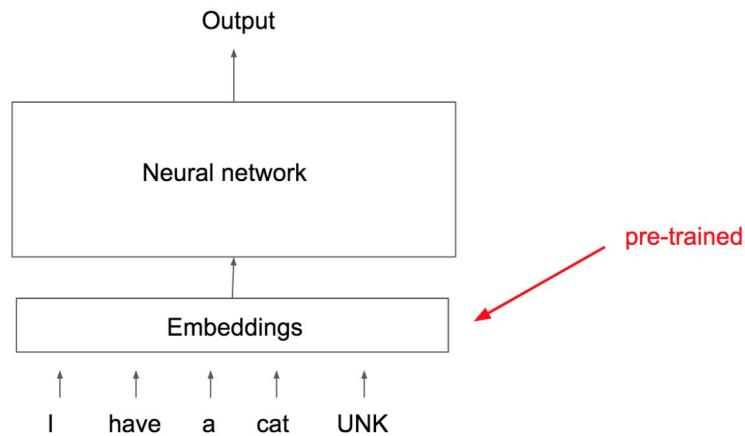


CBOW



Summary - Lecture 3

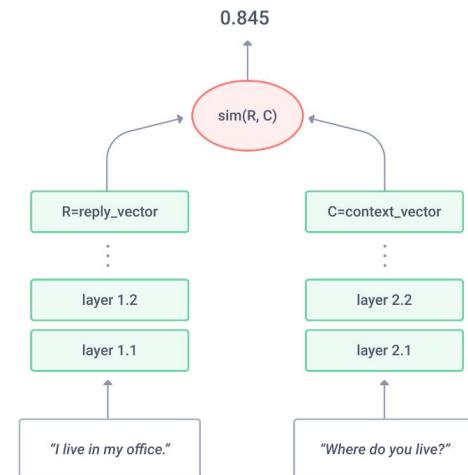
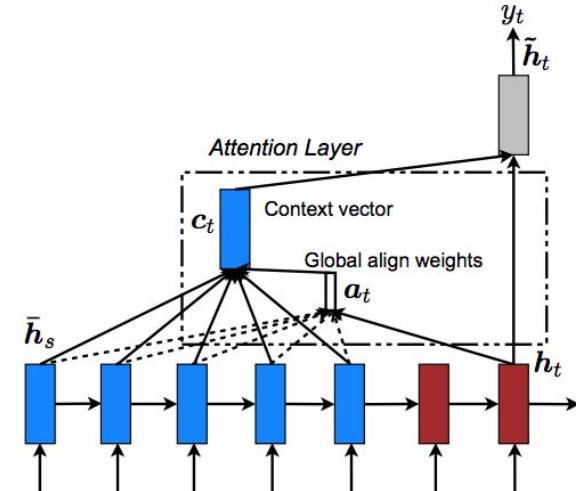
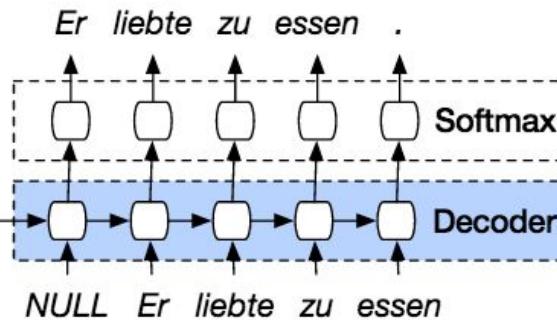
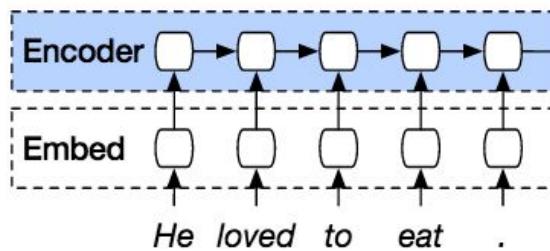
Recurrent neural networks (RNNs)



Summary - Lecture 4



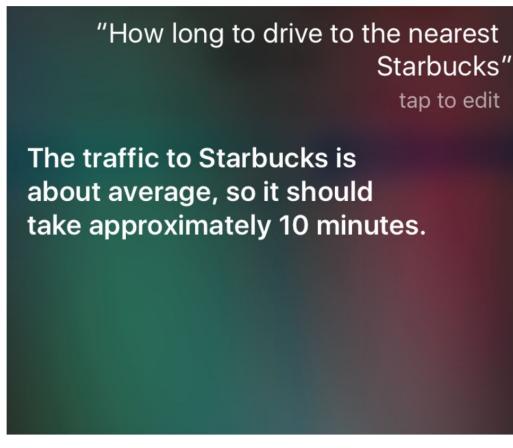
Figure 1: An example of NER application on an example text



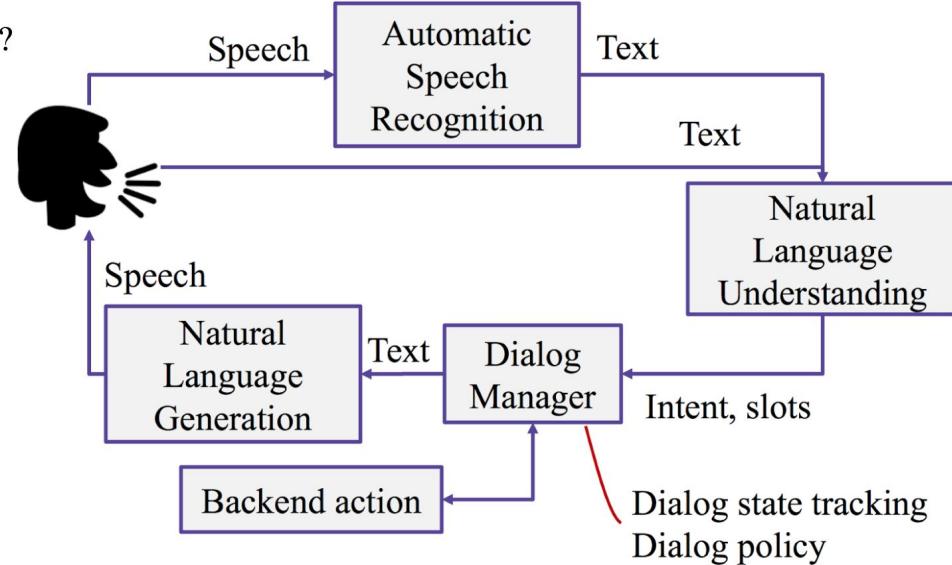
Summary - Lecture 4

Intent classification

- What does the user want?
- Which predefined scenario is the user trying to execute?



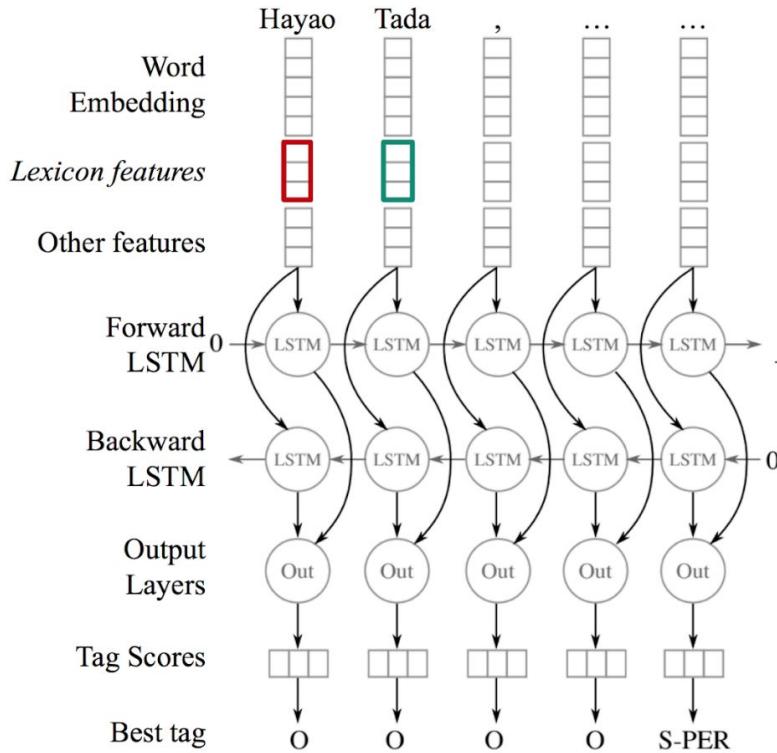
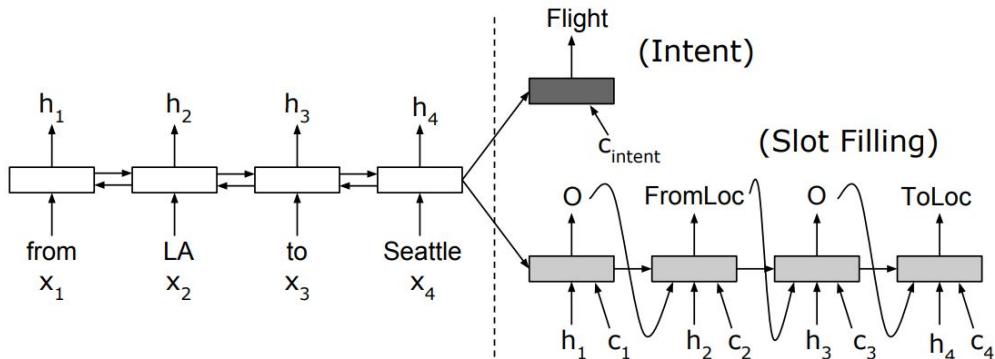
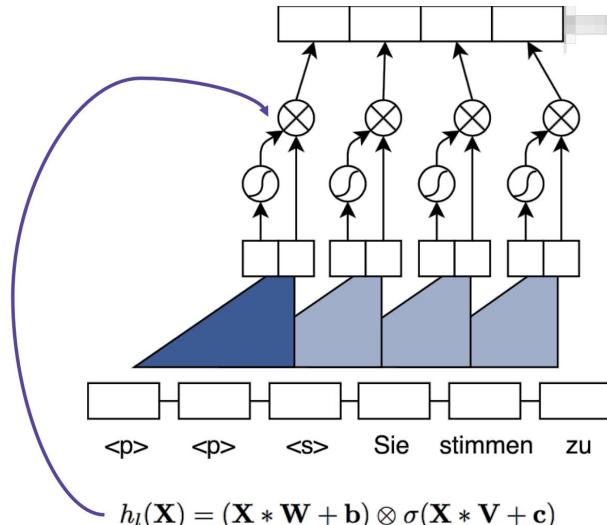
© Apple Siri



- Example of **BIO** scheme tagging:

Tokens	Show	me	the	way	to	History	Museum
Tags	O	O	O	O	O	B-TO	I-TO

Summary - Lecture 4



Links

- <https://www.toptal.com/machine-learning/clustering-algorithms>
- <https://web.stanford.edu/class/cs345a/slides/12-clustering.pdf>
- <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
- <https://medium.com/dailymotion/building-modern-recommender-systems-when-deep-learning-meets-product-principles-c79b16375109>
- <https://towardsdatascience.com/recommender-systems-with-deep-learning-architectures-1adf4eb0f7a6>
- <https://blog.statsbot.co/time-series-anomaly-detection-algorithms-1cef5519aef2>
- <https://arxiv.org/pdf/1707.07435.pdf>
- <https://blog.statsbot.co/recommendation-system-algorithms-ba67f39ac9a3>
- <https://static.googleusercontent.com/media/research.google.com/ru//pubs/archive/45530.pdf>
- <https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-datSc4-EMmixt.pdf>
- <https://arxiv.org/pdf/1902.07385.pdf>
- <https://medium.com/edureka/autoencoders-tutorial-cfdcebdefe37>
- <https://towardsdatascience.com/autoencoders-for-the-compression-of-stock-market-data-28e8c1a2da3e>