Кластеризация и EM-алгоритм

Полина Кириченко

Факультет компьютерных наук
Высшая школа экономики

Обучение с учителем и без учителя

X наблюдаемые переменные, а Y скрытые.
Цель: научиться предсказывать Y.

- Обучение с учителем (supervised learning)

- Обучение без учителя (unsupervised learning)

Обучение с учителем и без учителя

X наблюдаемые переменные, а Y скрытые.
Цель: научиться предсказывать Y.

- Обучение с учителем (supervised learning)
  Дан набор $< X, Y >$

- Обучение без учителя (unsupervised learning)
  Дан только набор $< X >$

Обучение с учителем и без учителя

X наблюдаемые переменные, а Y скрытые.
Цель: научиться предсказывать Y.

- Обучение с учителем (supervised learning)
  Дан набор $< X, Y >$
  - Легко сказать, правильный ли ответ выдаёт алгоритм

- Обучение без учителя (unsupervised learning)
  Дан только набор $< X >$
  - Непонятно, как определить хорошую работу алгоритма, нет “правильного ответа”

Обучение с учителем и без учителя

X наблюдаемые переменные, а Y скрытые.
Цель: научиться предсказывать Y.

- Обучение с учителем (supervised learning)
  Дан набор $< X, Y >$
  - Легко сказать, правильный ли ответ выдаёт алгоритм

- Обучение без учителя (unsupervised learning)
  Дан только набор $< X >$
  - Непонятно, как определить хорошую работу алгоритма, нет “правильного ответа”
  - Можем использовать результат не только для предсказания Y

Кластеризация

Дано:

- X – множество объектов
- Y – множество номеров кластеров
- $\rho(x, x')$ – функция расстояния между объектами
- обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$

Кластеризация

Дано:

- X – множество объектов
- Y – множество номеров кластеров
- $\rho(x, x')$ – функция расстояния между объектами
- обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$

Задача: разбить выборку на непересекающиеся подмножества, чтобы в смысле метрики ρ

- объекты одного кластера были как можно ближе
- объекты разных кластеров существенно отличались.

Неоднозначность решения

- Нет однозначно наилучшего критерия качества кластеризации
- Часто число кластеров неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием
- Результат кластеризации существенно зависит от метрики

K-means

Минимизация суммарного квадратичного отклонения точек кластеров от центров их кластеров:

$$SCORE = \sum_{i=1}^{K} \sum_{x \in C_i} ||x - c_i||^2$$

где C_i – кластеры, $c_i = \frac{1}{|C_k|} \sum_{x \in C_k} x$ – центры

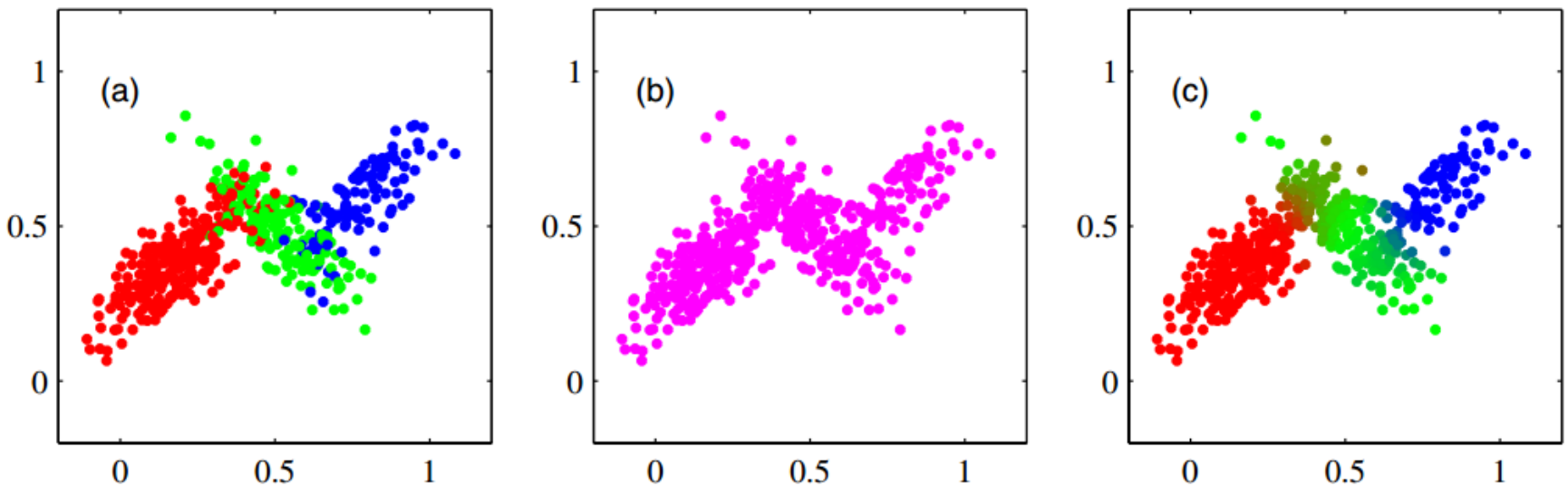
K-means

1. Выбрать K центров кластеров
2. Отенести каждую точку к ближайшему кластеру
3. Пересчитать центр каждого кластера
4. К шагу 2, до сходимости SCORE

На каждой итерации $O(kn)$ операций. Каждый шаг итерации сокращает $SCORE$.

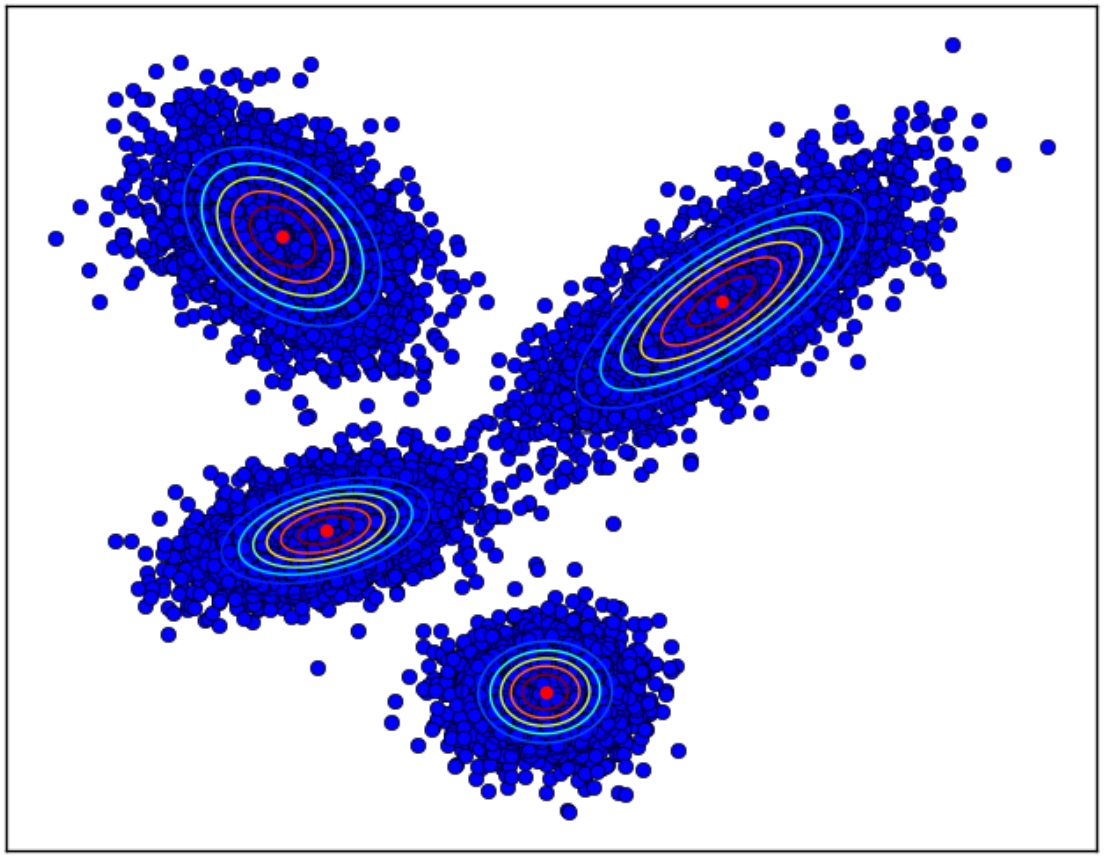
Clustering

- Hard clustering: each object belongs to a cluster or not
- Soft clustering: a likelihood of belonging to the cluster



Distribution models

Mixture of distributions $P_\theta(x) = \sum_i \pi_i p_{\theta_i}(x)$ where $\sum_i \pi_i = 1$ and $p_{\theta_i}(x)$ – individual pdf



Expectation-maximization algorithm

EM

Model $\mathscr{P} = \{P_\theta | \theta \in \Theta\}$
Data $X \sim \mathscr{L}(X) \in \mathscr{P}$
and latent Z with $P(X)$

EM for GMM

GMM $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$
$\theta = (\pi, \mu, \Sigma)$
$X \sim$ GMM, Z – gaussian index

Expectation-maximization algorithm

EM

Model $\mathscr{P} = \{P_\theta | \theta \in \Theta\}$
Data $X \sim \mathscr{L}(X) \in \mathscr{P}$
and latent Z with $P(X)$

Likelihood
$p(X|\theta) = \sum_Z p(X, Z|\theta) \to \max$
But Z are unknown!

EM for GMM

GMM $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$
$\theta = (\pi, \mu, \Sigma)$
$X \sim$ GMM, Z – gaussian index

Likelihood $\ln p(X, Z|\theta) =$
$= \sum_n \sum_k Z_{nk} \ln (\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k))$
where Z_{nk} an indicator $X_n \in Gaus_k$

Expectation-maximization algorithm

EM	EM for GMM
Model $\mathscr{P} = \{P_\theta \mid \theta \in \Theta\}$ Data $X \sim \mathscr{L}(X) \in \mathscr{P}$ and latent Z with $P(X)$	GMM $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$ $\theta = (\pi, \mu, \Sigma)$ $X \sim$ GMM, Z – gaussian index
Likelihood $p(X \mid \theta) = \sum_Z p(X, Z \mid \theta) \to \max$ But Z are unknown!	Likelihood $\ln p(X, Z \mid \theta) =$ $= \sum_n \sum_k Z_{nk} \ln \left(\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\right)$ where Z_{nk} an indicator $X_n \in Gaus_k$
Expectation $\mathbb{E}_Z \ln p(X, Z \mid \theta) =$ $= \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$ using posterior prob. $p(Z \mid X, \theta^{old})$	$\mathbb{E}_Z \ln p(X, Z \mid \pi, \mu, \Sigma) =$ $= \sum_n \sum_k \mathbb{E} Z_{nk} \left(\ln \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\right)$ where $\mathbb{E} Z_{nk} = P(Z_{nk} = 1 \mid X, \theta^{old})$

Expectation-maximization algorithm

EM	EM for GMM
Model $\mathscr{P} = \{P_\theta \mid \theta \in \Theta\}$ Data $X \sim \mathscr{L}(X) \in \mathscr{P}$ and latent Z with $P(X)$	GMM $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$ $\theta = (\pi, \mu, \Sigma)$ $X \sim$ GMM, Z – gaussian index
Likelihood $p(X \mid \theta) = \sum_Z p(X, Z \mid \theta) \to \max$ But Z are unknown!	Likelihood $\ln p(X, Z \mid \theta) =$ $= \sum_n \sum_k Z_{nk} \ln \left(\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\right)$ where Z_{nk} an indicator $X_n \in Gaus_k$
Expectation $\mathbb{E}_Z \ln p(X, Z \mid \theta) =$ $= \sum_Z p(Z \mid X, \theta^{old}) \ln p(X, Z \mid \theta)$ using posterior prob. $p(Z \mid X, \theta^{old})$	$\mathbb{E}_Z \ln p(X, Z \mid \pi, \mu, \Sigma) =$ $= \sum_n \sum_k \mathbb{E} Z_{nk} \left(\ln \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\right)$ where $\mathbb{E} Z_{nk} = P(Z_{nk} = 1 \mid X, \theta^{old})$

Maximization of $\mathbb{E}_Z \ln p(X, Z \mid \theta)$ with respsect to θ

EM-алгоритм в общем виде

- Инициалиация θ^{old}

EM-алгоритм в общем виде

- E-шаг. Вычислить апостериорные вероятности для скрытых переменных:

$$p(Z|X,\theta^{old}) = \frac{p(X,Z|\theta)}{p(X|\theta)} = \frac{p(X|Z,\theta)p(Z|\theta)}{\int p(X|Y,\theta)p(Y|\theta)dY}$$

(формула Байеса + формула полной вероятности)

Ожидаемое значение функции полного правдоподобия:

$$Q(\theta|\theta^{old}) = \sum_{Z} p(Z|X,\theta^{old}) \ln p(X,Z|\theta)$$

EM-алгоритм в общем виде

- M-шаг. Пересчёт параметров:

$$\theta^{new} = arg \max_{\theta} Q(\theta|\theta^{old})$$

Таким образом увеличивается ожидаемое правдоподобие, вычисляемое на E-шаге.

EM-алгоритм в общем виде

- Вычислить логарифм правдоподобия, выполнять алгоритм до сходимости

Доказательство

Улучшаем $Q(\theta|\theta^{old})$ вместо $\ln p(X|\theta)$, почему это работает?

KL-divergence

Kullback-Leibler divergence is a measure of the difference between two pdf P and Q (often P – prior, Q – estimated approximation of P).

- Generally : $D_{KL}(P||Q) = E_P\left[\ln \frac{P}{Q}\right]$
- Discrete: $D_{KL}(P||Q) = \sum_i p_i \ln \frac{p_i}{q_i}$
- Continuous: $D_{KL}(P||Q) = \int\limits_{\mathbb{R}} p(x) \ln \frac{p(x)}{q(x)} dx.$

$D_{KL}(P||Q) \geq 0$ with equality iff $P = Q$ alomost everywhere (without proof)

Доказательство

$$Q(\theta|\theta^{old}) = \sum_Z p(Z|X,\theta^{old}) \ln p(X,Z|\theta)$$

- $\ln p(X|\theta) = \mathcal{L}(q,\theta) + D_{KL}(q||p) \geq \mathcal{L}(q,\theta)$ where $q(Z)$ is distribution over Z
- E-step: $\mathcal{L}(q,\theta) \to \max_q$ with fixed $\theta^{old} \Rightarrow$ $\mathcal{L}(q,\theta) = Q(\theta|\theta^{old}) +$ const
- M-step: $\mathcal{L}(q,\theta) \to \max_\theta \Rightarrow Q(\theta|\theta^{old}) \to \max_\theta$

EM-алгоритм разделения смеси гауссиан

- Инициализация μ_k, Σ_k, π_k

EM-алгоритм разделения смеси гауссиан

- E-шаг. Вычислить апостериорные вероятности:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_i \pi_i N(x_n|\mu_i, \Sigma_i)}$$

π_k – априорная вероятность, γ_{nk} – апостериорная вероятность (“responsibility”)

EM-алгоритм разделения смеси гауссиан

- M-шаг. Пересчёт параметров:

$$N_k = \sum_{n=1}^{N} \gamma_{nk}, \quad \pi_k = \frac{N_k}{N},$$

$$\mu'_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n,$$

$$\Sigma'_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu'_k)(x_n - \mu'_k)^T$$

EM-алгоритм разделения смеси гауссиан

- Логарифм функции правдоподобия, к шагу 2

Incremental EM

A – mini-batch, update just the summands corresponding to data points from A

$$\pi_k^{new} = \frac{N_k^{new}}{N}$$

$$\mu_k^{new} = \mu_k^{old} + \frac{1}{N_k^{new}} \sum_{n \in A} (\gamma_{nk}^{new} - \gamma_{nk}^{old})(x_n - \mu_k^{old})$$

$$\Sigma_k^{new} = \Sigma_k^{old} + \frac{1}{N_k^{new}} \Big(\sum_{n \in A} (\gamma_{nk}^{new} - \gamma_{nk}^{old}) \Big((x_n - \mu_k^{new})(x_n - \mu_k^{new})^T - \Sigma_k^{old} \Big) + \\ + N_k^{old} (\mu_k^{new} - \mu_k^{old})(\mu_k^{new} - \mu_k^{old})^T \Big)$$

Stochastic optimization

With responsibilities γ_{nk} computed for mini-batch on E-step, M-step is:

$$\frac{\partial \ln p(X|\pi,\mu,\Sigma)}{\partial \mu_k} = \Sigma_k^{-1} \sum_{n \in A} \gamma_{nk}(x_n - \mu_k)$$

$$\frac{\partial \ln p(X|\pi,\mu,\Sigma)}{\partial \Sigma_k} = \sum_{n \in A} \frac{\gamma_{nk}}{2}(-\Sigma_k^{-1} + \Sigma_k^{-1}(x_n - \mu_k)(x_n - \mu_k)\Sigma_k^{-1})$$

$$\frac{\partial \ln p(X|\pi,\mu,\Sigma)}{\partial \pi_k} = \sum_{n \in A} \frac{\gamma_{nk}}{\pi_k}$$

Σ re-estimation formula contains matrix subtraction: to stay positive definite add $+\epsilon$ to diagonal

Learning rate

- $\frac{c}{t}$ or $\frac{c}{\sqrt{t}}$ where c is constant
- Momentum $\Delta\theta_{t+1} = \alpha\Delta\theta_t + \beta g_t$
- Adagrad $\Delta\theta_t = \frac{\alpha}{\sqrt{\sum_{\tau=1}^{t} g_\tau^2}} g_t$
- Adadelta

Эксперименты на модельных данных

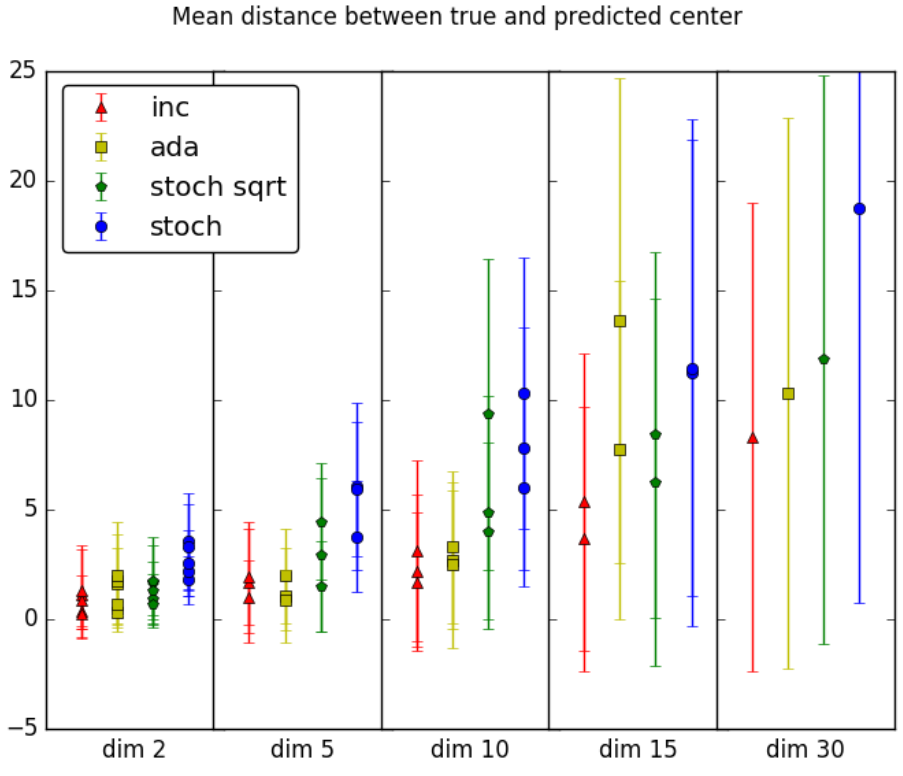


Рис.: Comparison of Stochastic Gradient EM with learning rates $\frac{c}{t}$ and $\frac{c}{\sqrt{t}}$, Adadelta and incremental EM. 2 epochs training with mini-batch size equal to 1. Constants c are chosen as showing the best quality in cross-validation, Adadelta hyperparameters are chosen as suggested in the original paper.

Эксперименты на модельных данных

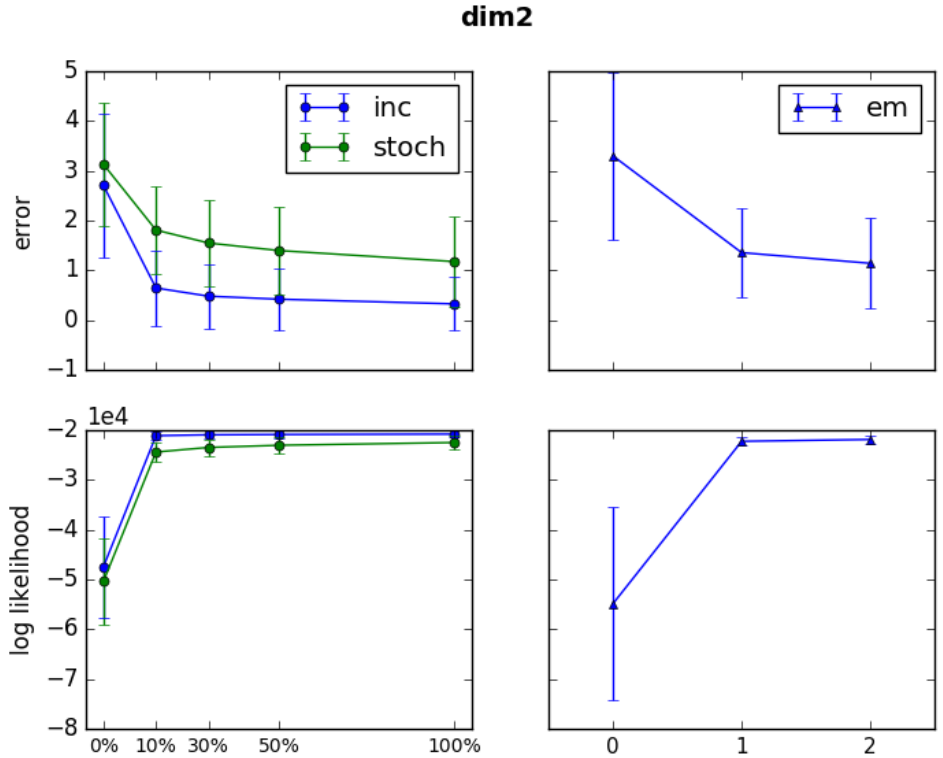
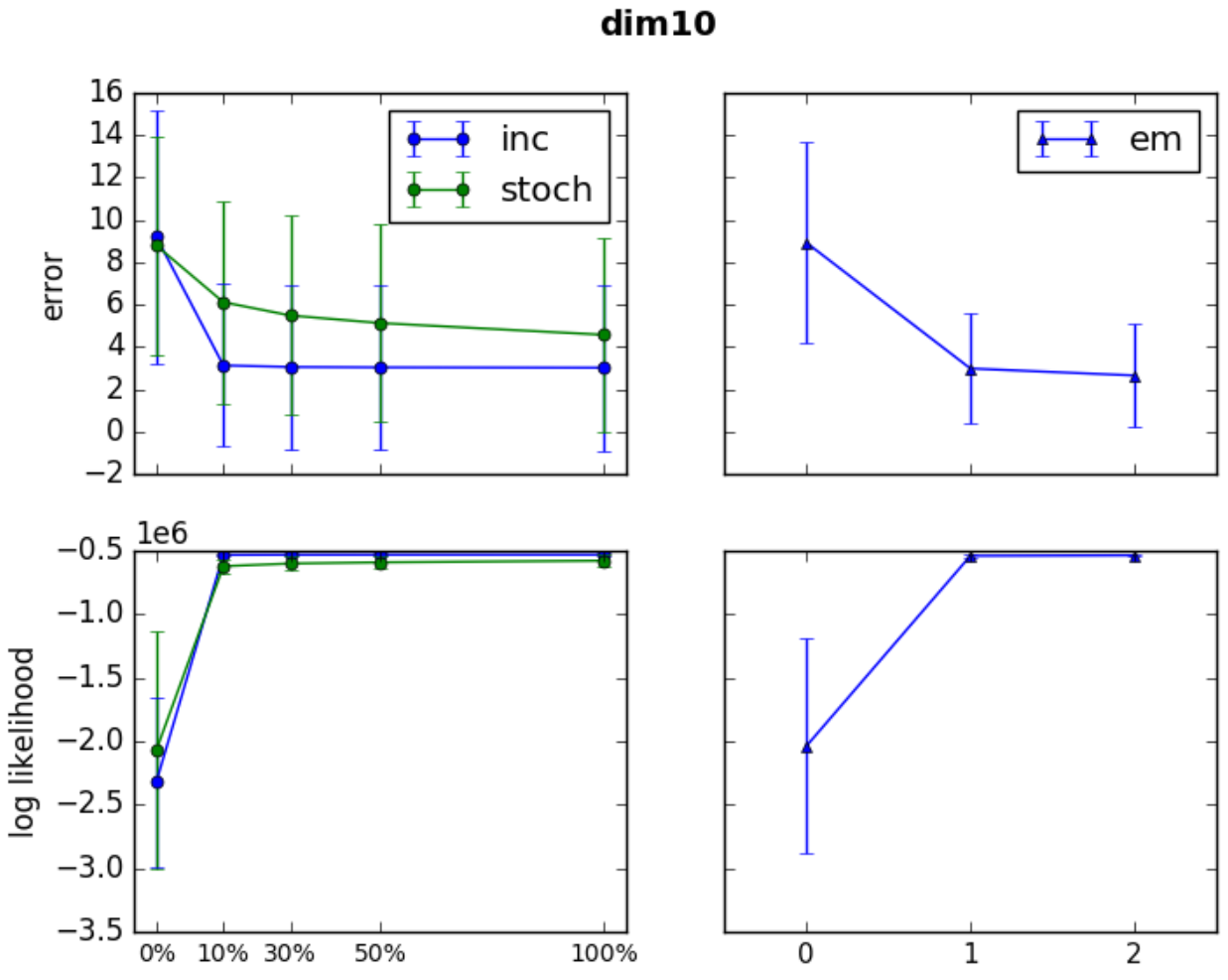


Рис.: Each set of 4 pictures corresponds to a particular dataset. Picture's left column: mean error and log likelihood after viewing 10, 30, 50 and 100% of the dataset for Stochastic Gradient EM with learning rate $\frac{0.05}{\sqrt{t}}$ and Incremental EM. Picture's right column: mean error and log likelihood after 1 and 2 iterations of classic EM.

Эксперименты на модельных данных



Выводы

- SG-EM с тепом обучения $\frac{c}{t}$ не сильно хуже «навороченных» методов оптимизации
- Инкрементный метод лучше SG-EM, но не для всякой модели будут формулы пересчёта

Спасибо за внимание!