

Введение в Байесовские методы

Часть I

Байесовский подход

Подход	Классический (частотный)	Байесовский
Случайность	объективная неопределенность	субъективное незнание
Величины	случайные, детерминированные	случайные (с той или иной степенью неопределенности)
Метод вывода	максимизация правдоподобия	теорема Байеса
Оценки	точечные, интервальные	апостериорное распределение
Применимость	$N \gg 1$	$\forall N$

Достоинства Байесовского подхода

- Регуляризация
- Композитность
- Гибкость
- Масштабируемость
- Выбор структурных параметров модели
- Latent Variable Modeling
- ...

Теорема Байеса

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}$$

Теорема Байеса

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}$$

Ищем

Теорема Байеса

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta}$$

Как посчитать?

Обзор методов Байесовского вывода

Рассмотрим модель $p(X, T, \theta)$.

Свойство модели	Метод вывода	Рассчитываем
Полное сопряжение	Полный Байесовский вывод	$p(T, \theta X)$
Условное сопряжение	MFA	$q(T) q(\theta)$
Сопряжение относительно T ($p(T X, \theta) \text{ — ОК}$)	EM ($p(\theta X) \rightarrow \max$)	$q(T) \delta(\theta - \theta^*)$
Сопряжение относительно θ	ME	$\delta(T - T^*) q(\theta)$
Условное сопряжение относительно T	Variational EM	$\prod_i q_i(T_i) \delta(\theta - \theta^*)$
Условное сопряжение относительно θ	Variational ME	$\delta(T - T^*) \prod_i q_i(\theta_i)$
Нет сопряжения	Poor Bayes (e. g. MAP, Crisp EM...)	$\delta(T - T^*) \delta(\theta - \theta^*)$

Байес для бедных или MAP-estimate

$$p(\theta \mid X) \approx \delta(\theta - \theta_*)$$

Байес для бедных или MAP-estimate

$$p(\theta \mid X) \approx \delta(\theta - \theta_*)$$

$$\begin{aligned}\theta_{MAP} &= \arg \min_{\theta_*} KL(\delta(\theta - \theta_*) \parallel p(\theta \mid X)) = \arg \max_{\theta} p(\theta \mid X) = \\ &= \arg \max_{\theta} \log p(\theta \mid X) = \arg \max_{\theta} [\log p(X \mid \theta) + \log p(\theta)]\end{aligned}$$

Байес для бедных или MAP-estimate

$$p(\theta \mid X) \approx \delta(\theta - \theta_*)$$

$$\begin{aligned}\theta_{MAP} &= \arg \min_{\theta_*} KL(\delta(\theta - \theta_*) \parallel p(\theta \mid X)) = \arg \max_{\theta} p(\theta \mid X) = \\ &= \arg \max_{\theta} \log p(\theta \mid X) = \arg \max_{\theta} [\log p(X \mid \theta) + \log p(\theta)]\end{aligned}$$

Можно не считать нормировочную константу!

Сопряженные распределения

Определение: пусть $p(x|\theta) \sim \Phi(\theta)$, $p(\theta|\alpha) \sim U(\alpha)$. Семейства распределений $\Phi(\theta)$ и $U(\alpha)$ называются *сопряженными*, если $p(\theta|X) \sim U(\alpha')$.

Сопряженные распределения

Определение: пусть $p(x|\theta) \sim \Phi(\theta)$, $p(\theta|\alpha) \sim U(\alpha)$. Семейства распределений $\Phi(\theta)$ и $U(\alpha)$ называются *сопряженными*, если $p(\theta|X) \sim U(\alpha')$.

Если имеется сопряжение, можно не считать нормировочный интеграл!

Сопряженные распределения

Определение: пусть $p(x | \theta) \sim \Phi(\theta)$, $p(\theta | \alpha) \sim U(\alpha)$. Семейства распределений $\Phi(\theta)$ и $U(\alpha)$ называются *сопряженными*, если $p(\theta | X) \sim U(\alpha')$.

Пример:

$$p(x | \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} \sim \mathcal{N}(x | \mu, 1)$$

$$p(\mu | X) = \frac{p(\mu) \prod_{i=1}^N p(x_i | \mu)}{p(X)} = \frac{p(\mu) \exp\left(-\frac{\sum_{i=1}^N (\mu - x_i)^2}{2}\right)}{\text{const}(\mu)}$$

$$p(\mu) — ?$$

Сопряженные распределения

Определение: пусть $p(x | \theta) \sim \Phi(\theta)$, $p(\theta | \alpha) \sim U(\alpha)$. Семейства распределений $\Phi(\theta)$ и $U(\alpha)$ называются *сопряженными*, если $p(\theta | X) \sim U(\alpha')$.

Пример:

$$p(x | \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} \sim \mathcal{N}(x | \mu, 1)$$

$$p(\mu | X) = \frac{p(\mu) \prod_{i=1}^N p(x_i | \mu)}{p(X)} = \frac{p(\mu) \exp\left(-\frac{\sum_{i=1}^N (\mu - x_i)^2}{2}\right)}{\text{const}(\mu)}$$

$$p(\mu) \sim \mathcal{N}(\mu | m, \sigma^2)$$

Экспоненциальный класс распределений

$$p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp \left(\theta^T u(x) \right)$$

$$f(x) \geq 0 \quad \forall u(x)$$

$$g(\theta) = \int f(x) \exp \left(\theta^T u(x) \right) dx$$

Экспоненциальный класс распределений

- Огромное количество семейств распределений (Нормальное, Гамма, Бета, Дирихле, Биномиальное, Пуассона...)
- $u(x)$ — достаточные статистики (MLE — функция достаточной статистики)
- Имеют место следующие соотношения:

$$\frac{\partial \log g(\theta)}{\partial \theta_j} = \mathbb{E}_{x \sim p(x|\theta)} u_j(x) \qquad \frac{\partial^2 \log g(\theta)}{\partial \theta_i \partial \theta_j} = \text{cov}(u_i(x), u_j(x))$$

- Логарифм правдоподобия — вогнутая функция
- ...

Сопряжение к ЭКР

$$p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp \left(\theta^T u(x) \right)$$

Сопряжение к ЭКР

$$p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp \left(\theta^T u(x) \right) \quad p(\theta \mid \eta, \nu) = \frac{\exp \left(\theta^T \eta \right)}{g^\nu(\theta) h(\eta, \nu)}$$

Сопряжение к ЭКР

$$p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp \left(\theta^T u(x) \right) \quad p(\theta \mid \eta, \nu) = \frac{\exp \left(\theta^T \eta \right)}{g^\nu(\theta) h(\eta, \nu)}$$

$$p(\theta \mid X) = \frac{\exp \left(\theta^T \eta' \right)}{g^{\nu'}(\theta) h(\eta', \nu')}$$

$$\eta' = \eta + \sum_{i=1}^N u(x_i) \quad \nu' = \nu + N$$

Байесовские модели в машинном обучении

X — наблюдаемые переменные, T — целевые переменные, θ — параметры.

Дискриминативная модель: $p(t, \theta \mid x) = p(t \mid x, \theta)p(\theta)$

Этап	Дано	Оцениваем	Считаем
Обучение	X_{tr}, T_{tr}	θ	$p(\theta \mid X_{tr}, T_{tr})$
Тестирование	x_*	t_*	$p(t_* \mid x_*, X_{tr}, T_{tr})$

$$p(\theta \mid X_{tr}, T_{tr}) = \frac{p(\theta) \prod_{i=1}^N p(t_i \mid x_i, \theta)}{p(T_{tr} \mid X_{tr})} = \frac{p(\theta) \prod_{i=1}^N p(t_i \mid x_i, \theta)}{\int p(\theta) \prod_{i=1}^N p(t_i \mid x_i, \theta) d\theta}$$

$$p(t_* \mid x_*, X_{tr}, T_{tr}) = \int p(t_* \mid x_*, \theta) p(\theta \mid X_{tr}, T_{tr}) d\theta$$

Байесовский выбор модели

Пусть есть несколько дискриминативных моделей:

$$p_j(t, \theta \mid x) = p_j(t \mid x, \theta)p_j(\theta), j \in J$$

Байесовский выбор модели

Пусть есть несколько дискриминативных моделей:

$$p_j(t, \theta \mid x) = p_j(t \mid x, \theta)p_j(\theta), j \in J$$

Принцип наибольшей обоснованности:

$$j^* = \arg \max_j p_j(T_{tr} \mid X_{tr}) = \arg \max_j \int p_j(\theta) \prod_{i=1}^N p_j(t_i \mid x_i, \theta) d\theta$$

Байесовский выбор модели

Пусть есть несколько дискриминативных моделей:

$$p_j(t, \theta \mid x) = p_j(t \mid x, \theta)p_j(\theta), j \in J$$

Принцип наибольшей обоснованности:

$$j^* = \arg \max_j p_j(T_{tr} \mid X_{tr}) = \arg \max_j \int p_j(\theta) \prod_{i=1}^N p_j(t_i \mid x_i, \theta) d\theta$$

Параметры модели настраиваются путем максимизации неполного правдоподобия (обоснованности), поэтому применение ПНО не приводит к переобучению.

Байесовская линейная регрессия

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \mathcal{N}(t \mid w^T x, \beta^{-1})\mathcal{N}(w \mid 0, \alpha^{-1}I)$$

Байесовская линейная регрессия

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \mathcal{N}(t \mid w^T x, \beta^{-1})\mathcal{N}(w \mid 0, \alpha^{-1}I)$$

Обучение

Пусть (X, T) — обучающая выборка.

Получение апостериорного распределения — легкая задача, так как имеем сопряжение.

$$p(w \mid X, T) = \frac{p(T \mid X, w)p(w)}{\int p(T \mid X, w)p(w)dw} \sim \mathcal{N}(w \mid w_{MP}, \Sigma)$$

$$\Sigma = (\beta X^T X + \alpha I)^{-1} \quad w_{MP} = \beta \Sigma X^T T$$

Байесовская линейная регрессия

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \mathcal{N}(t \mid w^T x, \beta^{-1})\mathcal{N}(w \mid 0, \alpha^{-1}I)$$

Тестирование

Пусть (X, T) — обучающая выборка. x^* — новый объект.

Получение распределения на прогноз имеет ту же сложность (в смысле аналитического вывода), что и подсчет апостериорного распределения.

$$p(t_* \mid x_*, X, T) = \int p(t_* \mid x_*, w)p(w \mid X, T)dw \sim \mathcal{N}(t_* \mid y_*, \sigma_*^2)$$

$$y_* = w_{MP}^T x_*$$

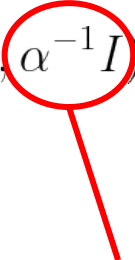
$$\sigma_*^2 = \beta^{-1} + x_*^T \Sigma x_*$$

Что можно улучшить?

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \mathcal{N}(t \mid w^T x, \beta^{-1})\mathcal{N}(w \mid 0, \alpha^{-1}I)$$

Что можно улучшить?

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \mathcal{N}(t \mid w^T x, \beta^{-1})\mathcal{N}(w \mid 0, \alpha^{-1}I)$$



Стандартная L2-
регуляризация

Relevance Vector Regression

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \mathcal{N}(t \mid w^T x, \beta^{-1})\mathcal{N}(w \mid 0, A^{-1})$$

$A = \text{diag}\{\alpha_1, \dots, \alpha_d\}$ — получаем избирательную регуляризацию.

RVR: подбор модели

Как подобрать гиперпараметры A и β (параметры регуляризации)? ПНО!

$$\int Q(w \mid A, \beta) dw = Q(w_{MP} \mid A, \beta) \sqrt{(2\pi)^d \det \Sigma} \rightarrow \max_{A, \beta}$$

$$Q(w \mid A, \beta) = p(T \mid X, w, \beta) p(w \mid A)$$

$$\Sigma = (\beta X^T X + A)^{-1} \quad w_{MP} = \beta \Sigma X^T T$$

w_{MP} зависит от A и β , поэтому приходится прибегать к итерационной оптимизации.

Variational lower bound

$$f(x) \rightarrow \max_x$$

$$g(x, \xi) : (\forall x, \xi \ f(x) \geq g(x, \xi)) \wedge (\forall x \exists \xi_x \ f(x) = g(x, \xi_x))$$

Определим итерационный процесс оптимизации $f(x)$:

$$\begin{cases} x_n = \arg \max_x g(x, \xi_{n-1}) \\ \xi_n = \arg \max_{\xi} g(x_n, \xi) \end{cases}$$

RVR: Variational lower bound

Здесь:

$$x = (A, \beta) \quad \xi = w$$

$$f(A, \beta) = Q(w_{MP} \mid A, \beta) \sqrt{(2\pi)^d \det \Sigma}$$

$$g(w, A, \beta) = Q(w \mid A, \beta) \sqrt{(2\pi)^d \det \Sigma}$$

RVR: алгоритм обучения

Вход: Обучающая выборка $\{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$;
Матрица обобщенных признаков $\Phi = \{\phi_j(\mathbf{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов \mathbf{w} , матрица Σ и оценка дисперсии шума β^{-1} для решающего правила $t_*(\mathbf{x}) = \sum_{j=1}^m w_j \phi_j(\mathbf{x})$, $\sigma_*^2(\mathbf{x}) = \beta^{-1} + \boldsymbol{\phi}^T(\mathbf{x}_*) \Sigma \boldsymbol{\phi}(\mathbf{x}_*)$;

1: инициализация: $\alpha_i := 1$, $i = 1, \dots, m$, $\beta := 1$, AlphaBound := 10^{12} , WeightBound := 10^{-6} , NumberOfIterations := 100;

2: **для** $k = 1, \dots, \text{NumberOfIterations}$

3: $A := \text{diag}(\alpha_1, \dots, \alpha_m)$;

4: $\Sigma := (\beta \Phi^T \Phi + A)^{-1}$;

5: $\mathbf{w}_{MP} := \Sigma \beta \Phi^T \mathbf{t}$;

6: **для** $j = 1, \dots, m$

7: **если** $w_{MP,j} < \text{WeightBound}$ или $\alpha_j > \text{AlphaBound}$ **то**

8: $w_{MP,j} := 0$, $\alpha_j := +\infty$, $\gamma_j := 0$;

9: **иначе**

10: $\gamma_j := 1 - \alpha_j \Sigma_{jj}$, $\alpha_j := \frac{\gamma_j}{w_{MP,j}^2}$;

11: $\beta := \frac{n - \sum_{j=1}^m \gamma_j}{\|\mathbf{t} - \Phi \mathbf{w}_{MP}\|^2}$

RVR vs L1 vs L2

Здесь решается модельная задача: зашумленным полиномом третьей степени сгенерированы данные для задачи регрессии. Нужно на этих данных обучить многочлен степени, не превышающей 20. Предлагается сравнить три модели: гребневую регрессию, L1-регрессию (Lasso) и RVR, и сравнить ошибку на тестовой выборке и качество отобранных признаков.

Relevance Vector Regression

Features remaining: 3 / 21

Train error: 9052.79670931

Test error: 9793.46412659

Ridge Regression

Features remaining: NA (no sparsity)

Train error: 8090.9980939

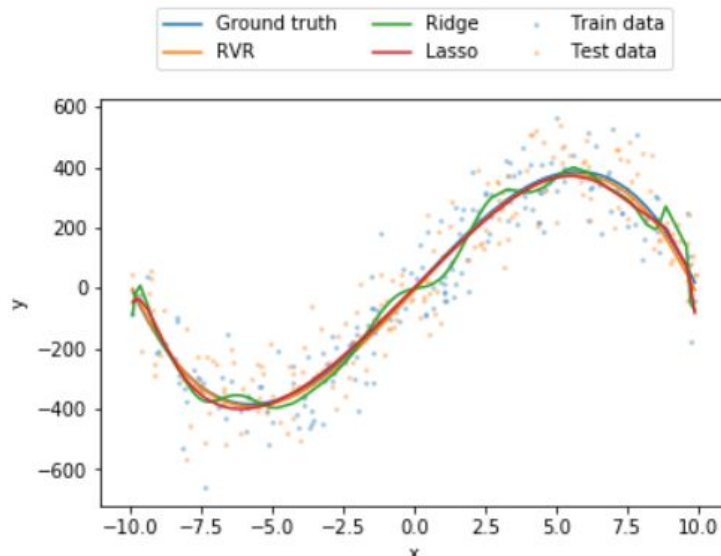
Test error: 12255.2357566

Lasso Regression

Features remaining: 19 / 21

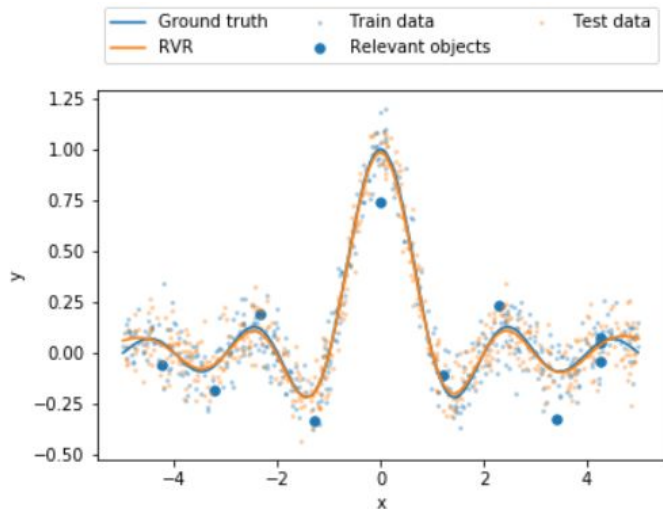
Train error: 8941.22847976

Test error: 10237.8405027



RVR vs SVR vs L1

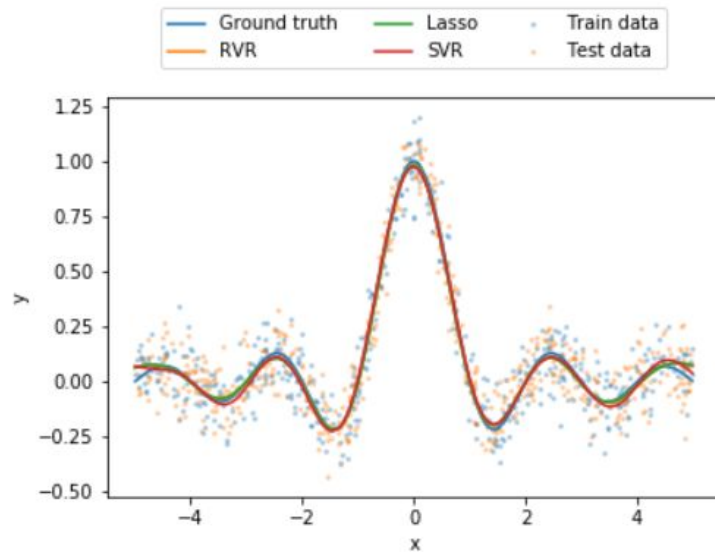
Здесь решается другая модельная задача: необходимо восстановить зашумленную функцию $\text{sinc}(x)$. Предлагается применить kernel trick с RBF-ядром, обучить три модели: SVM-регрессию (SVR), L1-регрессию (Lasso) и RVR, и сравнить ошибку на тестовой выборке и качество отобранных опорных / релевантных объектов.



Relevance Vector Regression
Objects remaining: 11 / 500
Train error: 0.00944546665494
Test error: 0.00930529012434

Lasso Regression
Objects remaining: 148 / 500
Train error: 0.00945624997543
Test error: 0.00934661882521

Support Vector Regression
Objects remaining: 163 / 500
Train error: 0.00947339517678
Test error: 0.00948180161155



Кратко о RVC

Для решения задачи классификации используется следующая дискриминативная модель:

$$p(t, w \mid x) = p(t \mid x, w)p(w) = \frac{1}{1 + \exp(-tw^T x)} \mathcal{N}(w \mid 0, A^{-1})$$

Сопряжения больше нет, поэтому приходится прибегать к различным приближениям интегралов (например, приближение Лапласа).

RVC: алгоритм обучения

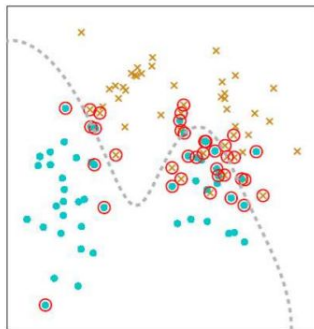
Вход: Обучающая выборка $\{\mathbf{x}_i, t_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$;

Матрица обобщенных признаков $\Phi = \{\phi_j(\mathbf{x}_i)\}_{i,j=1}^{n,m}$;

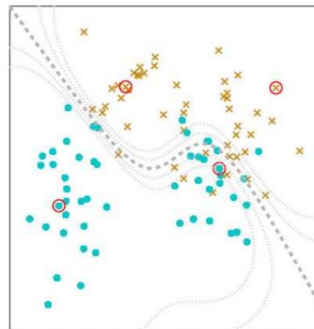
Выход: Набор весов \mathbf{w} для решающего правила $t_*(\mathbf{x}) = \sum_{j=1}^m w_j \phi_j(\mathbf{x})$;

- 1: инициализация: $\alpha_i := 1$, $i = 1, \dots, m$, $\mathbf{w}_{MP} = \mathbf{t}$, AlphaBound := 10^{12} , WeightBound := 10^{-6} , NumberOfIterations := 100;
- 2: **для** $k = 1, \dots, \text{NumberOfIterations}$
- 3: $A := \text{diag}(\alpha_1, \dots, \alpha_m)$;
- 4: **повторять**
- 5: **для** $i = 1, \dots, n$
- 6: $s_i := 1 / (1 + \exp(t_i \sum_{j=1}^m w_{MP,j} \phi_j(\mathbf{x}_i)))$;
- 7: $R := \text{diag}(s_1(1 - s_1), \dots, s_n(1 - s_n))$;
- 8: $\mathbf{z} := \Phi \mathbf{w}_{MP} + R^{-1}(\mathbf{s} - \mathbf{t})$;
- 9: $\Sigma := (\Phi^T R \Phi + A)^{-1}$;
- 10: $\mathbf{w}_{MP} := \Sigma \Phi^T R \mathbf{z}$;
- 11: **пока** $\|\mathbf{w}_{MP}^{new} - \mathbf{w}_{MP}^{old}\|$ меняется больше, чем на заданную величину
- 12: **для** $j = 1, \dots, m$
- 13: **если** $w_{MP,j} < \text{WeightBound}$ или $\alpha_j > \text{AlphaBound}$ **то**
- 14: $w_{MP,j} := 0$, $\alpha_j := +\infty$, $\gamma_j := 0$;
- 15: **иначе**
- 16: $\alpha_j := \frac{1 - \alpha_j \Sigma_{jj}}{w_{MP,j}^2}$;

RVC vs SVM



SVM



RVM

- SVM:

$$\sum_{i=1}^n [1 - t_i f(\mathbf{x}_i, \mathbf{w})]_+ + \gamma \sum_{j=1}^m w_j^2 \rightarrow \min_{\mathbf{w}}$$

- RVM:

$$\sum_{i=1}^n \log(1 + \exp(-t_i f(\mathbf{x}_i, \mathbf{w}))) + \sum_{j=1}^m \alpha_j w_j^2 \rightarrow \min_{\mathbf{w}}$$

ЕМ-алгоритм

Что хотим?

Получить MLE параметров θ , промаксимизировав неполное правдоподобие:

$$p(X \mid \theta) = \prod_{i=1}^N p(x_i \mid \theta) \rightarrow \max_{\theta}$$

ЕМ-алгоритм

Что мешает?

Случай 1. $p(X | \theta)$ не лежит в экспоненциальном классе, но если ввести некоторые “дополнительные” (“латентные” или “скрытые”) переменные T , то $p(X, T | \theta)$ лежит в ЭКР \Rightarrow легко оптимизируется.

ЕМ-алгоритм

Что мешает?

Случай 2. Наблюдаем лишь некоторые переменные (X), а часть переменных — скрыта (T). Хотим промаксимизировать неполное правдоподобие $p(X \mid \theta)$, но модель $p(X, T \mid \theta)$ более естественна.

ЕМ-алгоритм

Что мешает?

Случай 3. Пусть задана вероятностная модель $p(X, T \mid \theta)$. Тогда:

$$p(X \mid \theta) = \prod_{i=1}^N p(x_i \mid \theta) = \prod_{i=1}^N \int p(x_i, T \mid \theta) dT$$

Интегралы могут и не браться \Rightarrow хотим оптимизировать функцию, которую не можем рассчитать!

ЕМ-алгоритм

Выход есть? **Да!**

ЕМ-алгоритм

Пусть задано совместное распределение $p(X, T|\Theta)$.

X – набор наблюдаемых переменных,

T – набор ненаблюдаемых переменных,

Θ – набор параметров модели.

$$\log p(X|\Theta) = \log \int p(X, T|\Theta) dT \rightarrow \max_{\Theta}$$

ЕМ-алгоритм

$$\begin{aligned}\log p(X|\Theta) &= \int q(T) \log p(X|\Theta) dT = \int q(T) \log \frac{p(X, T|\Theta)}{p(T|X, \Theta)} dT = \int q(T) \log \left[\frac{p(X, T|\Theta)}{q(T)} \frac{q(T)}{p(T|X, \Theta)} \right] dT = \\ &= \underbrace{\int q(T) \log p(X, T|\Theta) dT}_{\mathcal{L}(q)} - \underbrace{\int q(T) \log q(T) dT + \int q(T) \log \frac{p(T|X, \Theta)}{q(T)} dT}_{\text{KL}(q||p(T|X, \Theta))}.\end{aligned}$$

ЕМ-алгоритм

$$\begin{aligned}\log p(X|\Theta) &= \int q(T) \log p(X|\Theta) dT = \int q(T) \log \frac{p(X, T|\Theta)}{p(T|X, \Theta)} dT = \int q(T) \log \left[\frac{p(X, T|\Theta)}{q(T)} \frac{q(T)}{p(T|X, \Theta)} \right] dT = \\ &= \underbrace{\int q(T) \log p(X, T|\Theta) dT}_{\mathcal{L}(q)} - \underbrace{\int q(T) \log q(T) dT + \int q(T) \log \frac{p(T|X, \Theta)}{q(T)} dT}_{\text{KL}(q||p(T|X, \Theta))}.\end{aligned}$$

$\log p(X|\Theta) \geq \mathcal{L}(q)$ — Нижняя граница
неполного правдоподобия

ЕМ-алгоритм

$$\begin{aligned}\log p(X|\Theta) &= \int q(T) \log p(X|\Theta) dT = \int q(T) \log \frac{p(X, T|\Theta)}{p(T|X, \Theta)} dT = \int q(T) \log \left[\frac{p(X, T|\Theta)}{q(T)} \frac{q(T)}{p(T|X, \Theta)} \right] dT = \\ &= \underbrace{\int q(T) \log p(X, T|\Theta) dT}_{\mathcal{L}(q)} - \underbrace{\int q(T) \log q(T) dT + \int q(T) \log \frac{p(T|X, \Theta)}{q(T)} dT}_{\text{KL}(q||p(T|X, \Theta))}.\end{aligned}$$

$\log p(X|\Theta) \geq \mathcal{L}(q)$ — Variational lower bound

$q(T) = p(T|X, \Theta)$ — Нижняя граница становится точной

E-шаг

$$q(T) = p(T|X, \Theta_{old}) = \frac{p(X, T|\Theta_{old})}{\int p(X, T|\Theta_{old})dT}$$

M-шаг

$$\mathbb{E}_{T|X, \Theta_{old}} \log p(X, T|\Theta) \rightarrow \max_{\Theta}$$

ЕМ-алгоритм: иллюстрация

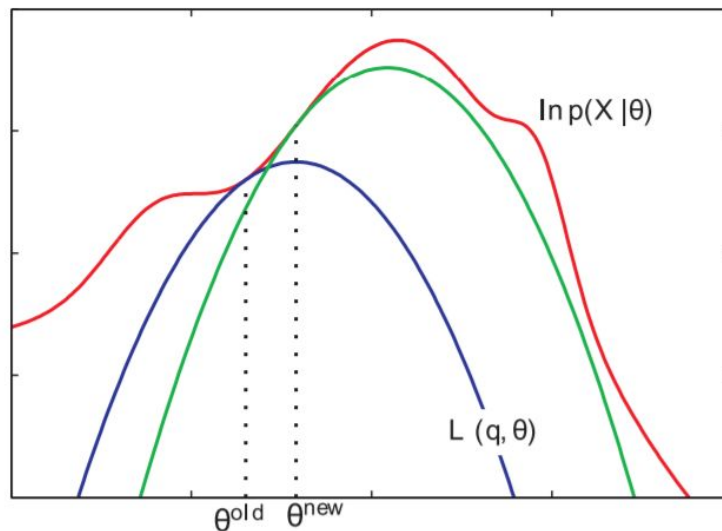


Рис. 2: Иллюстрация итерационного процесса в ЕМ-алгоритме. Нижняя оценка (3) обозначена через $L(q, \theta)$.

MAP-EM

Небольшое обобщение. Пусть задано совместное распределение:

$$p(X, T, \theta) = p(X, T \mid \theta)p(\theta)$$

$$p(\theta \mid X) \rightarrow \max_{\theta} \text{ — получаем MAP-оценку.}$$

Е-шаг не меняется.

$$\text{М-шаг: } \mathbb{E}_{T \sim q(T)} [\log p(X, T \mid \theta) + \log p(\theta)] \rightarrow \max_{\theta}$$

PCA

Principal component analysis (PCA) — один из базовых методов линейного снижения размерности в данных ($D \rightarrow d$).

- Сложность: $O(ND^2 + D^3 + N^3)$
- Не применим к данным с пропусками
- Проекция на единственное подпространство
- Ручной подбор размерности подпространства

PCA

Введем следующую вероятностную модель PCA:

$$p(\boldsymbol{x}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{t} + \boldsymbol{\mu}, \sigma^2 I), \quad p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t}|\mathbf{0}, I).$$

$\boldsymbol{W} \in \mathbb{R}^{D \times d}$ задает направляющие вектора гиперплоскости,
 $\boldsymbol{\mu} \in \mathbb{R}^D$ – смещение гиперплоскости относительно начала координат,
 $\sigma > 0$ определяет дисперсию шума в данных относительно гиперплоскости

PCA

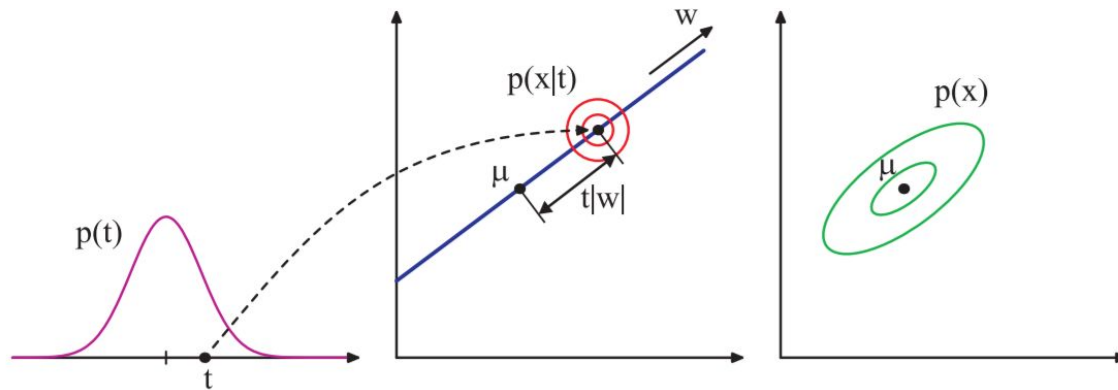


Рис. 3: Иллюстрация процесса генерации объекта в вероятностной модели PCA для $D = 2$ и $d = 1$. Наблюдаемое значение \mathbf{x} образуется путем генерирования значения скрытой компоненты t из априорного распределения $p(t)$ и последующего генерирования значения \mathbf{x} из изотропного нормального распределения с центром $\boldsymbol{\mu} + t\mathbf{w}$ и матрицей ковариации $\sigma^2 I$. Зеленые эллипсы показывают линии уровня плотности маргинального распределения $p(\mathbf{x})$.

РСА

Пусть объекты в выборке X независимы.

Приходим к следующему совместному распределению:

$$p(X, T|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{t}_n|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^N \mathcal{N}(\mathbf{x}_n|W\mathbf{t}_n + \boldsymbol{\mu}, \sigma^2 I) \mathcal{N}(\mathbf{t}_n|\mathbf{0}, I).$$

Решаем следующую задачу:

$$p(X|W, \boldsymbol{\mu}, \sigma) = \prod_{n=1}^N p(\mathbf{x}_n|W, \boldsymbol{\mu}, \sigma) \rightarrow \max_{W, \boldsymbol{\mu}, \sigma}.$$

PCA

Задача может быть решена в явном виде (действительно, это просто PCA):

$$\begin{aligned} p(\mathbf{x}_n|W, \boldsymbol{\mu}, \sigma) &= \int p(\mathbf{x}_n|\mathbf{t}_n, W, \boldsymbol{\mu}, \sigma)p(\mathbf{t}_n)d\mathbf{t}_n = \\ &= \int \mathcal{N}(\mathbf{x}_n|W\mathbf{t}_n + \boldsymbol{\mu}, \sigma^2 I)\mathcal{N}(\mathbf{t}_n|\mathbf{0}, I)d\mathbf{t}_n = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \sigma^2 I + WW^T). \end{aligned}$$

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \quad S = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})^T, \\ W &= Q(\Lambda - \sigma^2 I)^{1/2} R, \\ \sigma^2 &= \frac{1}{D-d} \sum_{i=d+1}^D \lambda_i. \end{aligned} \tag{3}$$

Здесь $Q = (\mathbf{q}_1 | \dots | \mathbf{q}_d) \in \mathbb{R}^{D \times d}$, $\mathbf{q}_1, \dots, \mathbf{q}_d$ – нормированные собственные вектора выборочной матрицы ковариации S , отвечающие наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$, $\|\mathbf{q}_i\| = 1$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$, R – произвольная ортогональная матрица размера $d \times d$.

РСА

Зная параметры $W, \boldsymbol{\mu}, \sigma$, задача поиска для объекта \boldsymbol{x} представления \boldsymbol{t} в пространстве \mathbb{R}^d сводится к вычислению математического ожидания условного распределения

$$p(\boldsymbol{t}|\boldsymbol{x}) = \frac{p(\boldsymbol{t}, \boldsymbol{x})}{\int p(\boldsymbol{t}, \boldsymbol{x}) d\boldsymbol{x}} = \mathcal{N}(\boldsymbol{t} | (\sigma^2 I + W^T W)^{-1} W^T (\boldsymbol{x} - \boldsymbol{\mu}), I + \sigma^{-2} W^T W),$$
$$\mathbb{E}_{\boldsymbol{t}|\boldsymbol{x}} \boldsymbol{t} = (\sigma^2 I + W^T W)^{-1} W^T (\boldsymbol{x} - \boldsymbol{\mu}).$$

Но что, если...

PCA + EM = ♥

E-шаг:

$$p(T|X, W_{old}, \sigma_{old}^2, \boldsymbol{\mu}_{old}) = \prod_{n=1}^N p(\mathbf{t}_n | \mathbf{x}_n, W_{old}, \sigma_{old}^2, \boldsymbol{\mu}_{old}),$$

$$p(\mathbf{t}_n | \mathbf{x}_n, W_{old}, \sigma_{old}^2, \boldsymbol{\mu}_{old}) = \mathcal{N}(\mathbf{t}_n | \boldsymbol{\mu}_n, \Sigma_n),$$

$$\boldsymbol{\mu}_n = M_{old} W_{old}^T (\mathbf{x}_n - \boldsymbol{\mu}_{old}),$$

$$\Sigma_n = \sigma_{old}^2 M_{old},$$

$$M_{old} = (W_{old}^T W_{old} + \sigma_{old}^2 I)^{-1}.$$

PCA + EM = ♥

M-шаг:

$$\begin{aligned}\boldsymbol{\mu}_{new} &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n, \\ W_{new} &= \left(\sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{new}) \mathbb{E} \mathbf{t}_n^T \right) \left(\sum_{n=1}^N \mathbb{E} \mathbf{t}_n \mathbf{t}_n^T \right)^{-1}, \\ \sigma_{new}^2 &= \frac{1}{ND} \sum_{n=1}^N \left((\mathbf{x}_n - \boldsymbol{\mu}_{new})^T (\mathbf{x}_n - \boldsymbol{\mu}_{new}) - 2 \mathbb{E} \mathbf{t}_n^T W_{new}^T (\mathbf{x}_n - \boldsymbol{\mu}_{new}) + \text{tr} W_{new}^T W_{new} \mathbb{E} \mathbf{t}_n \mathbf{t}_n^T \right).\end{aligned}\tag{5}$$

При этом необходимые статистики вычисляются следующим образом:

$$\begin{aligned}\mathbb{E} \mathbf{t}_n &= \boldsymbol{\mu}_n, \\ \mathbb{E} \mathbf{t}_n \mathbf{t}_n^T &= \Sigma_n + \boldsymbol{\mu}_n \boldsymbol{\mu}_n^T.\end{aligned}$$

PCA + EM = ♥

И что? А вот что:

Было: $O(ND^2 + D^3 + N^3)$

Стало: $O(i(NDd + d^3))$

Более того...

RSA + EM = учет пропусков в данных

Обозначим через K_n множество номеров известных значений признаков для объекта \mathbf{x}_n и U_n — множество пропущенных значений признаков для объекта \mathbf{x}_n , $K_n \cup U_n = \{1, \dots, D\}$. Соответственно $W_{K_n} = \{w_{ij}\}_{i \in K_n, j \in \{1, \dots, d\}}$. Вероятностная модель RSA с пропусками в данных выглядит следующим образом:

$$p(X_K, X_U, T | W, \sigma^2, \boldsymbol{\mu}) = \prod_{n=1}^N p(\mathbf{x}_{n,K_n}, \mathbf{x}_{n,U_n} | \mathbf{t}_n, W, \boldsymbol{\mu}, \sigma^2) p(\mathbf{t}_n),$$

$$p(\mathbf{x}_{n,K_n}, \mathbf{x}_{n,U_n} | \mathbf{t}_n, W, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}((\mathbf{x}_{n,K_n}, \mathbf{x}_{n,U_n}) | (W_{K_n} \mathbf{t}_n + \boldsymbol{\mu}_{K_n}, W_{U_n} \mathbf{t}_n + \boldsymbol{\mu}_{U_n}); \sigma^2 I),$$
$$p(\mathbf{t}_n) = \mathcal{N}(\mathbf{t}_n | \mathbf{0}, I).$$

PCA + EM = учет пропусков в данных

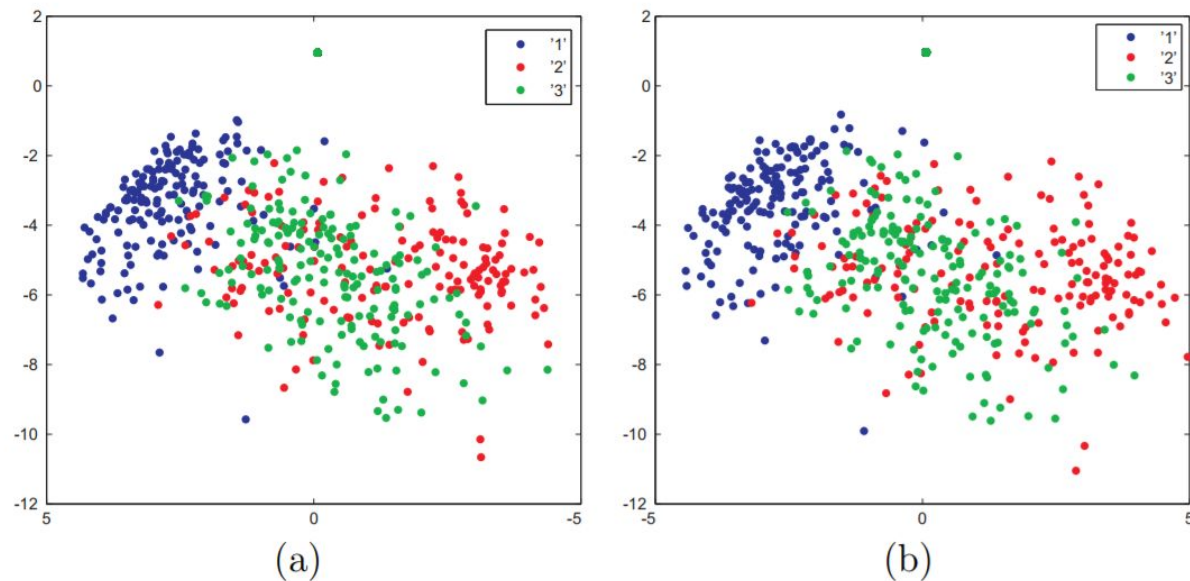


Рис. 4: Проекция выборки изображений цифр '1', '2', '3' на первые две главные компоненты для полных данных (а) и для выборки, в которой 30% случайно выбранных значений считаются пропущенными.

PCA + EM = BPCA

$$p(W|\boldsymbol{\alpha}) = \prod_{i=1}^D \sqrt{\frac{\alpha_i}{2\pi}} \exp\left(-\frac{\alpha_i}{2} \mathbf{w}_i^T \mathbf{w}\right).$$

$$p(X, T, W|\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}) = p(X|T, W, \boldsymbol{\mu}, \sigma)p(T)p(W|\boldsymbol{\alpha}).$$

$$p(X|\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}) = \int p(X|W, \boldsymbol{\mu}, \sigma)p(W|\boldsymbol{\alpha})dW \rightarrow \max_{\boldsymbol{\mu}, \sigma, \boldsymbol{\alpha}}.$$

PCA + EM = BPCA

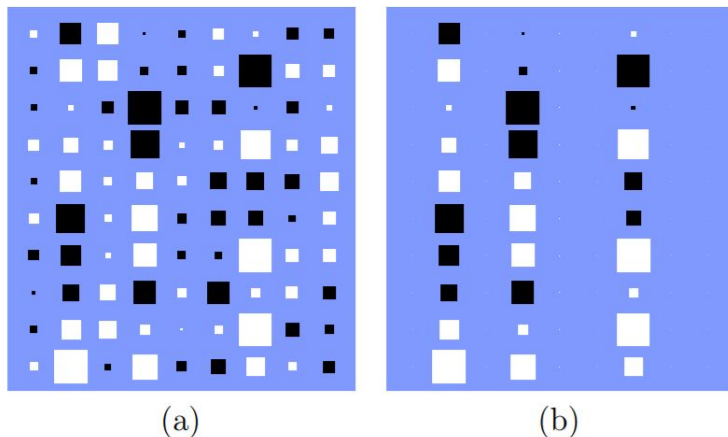


Рис. 5: Иллюстрация байесовского метода главных компонент. Модельные данные состоят из 300 объектов, сгенерированных из нормального распределения в пространстве размерности $D = 10$. При этом данные имеют стандартное отклонение 1 по трем направлениям в этом пространстве и стандартное отклонение 0.5 по остальным семи направлениям. На рис. а показана матрица W , полученная с помощью стандартного метода главных компонент (белые и черные квадраты соответствуют положительным и отрицательным значениям, величина квадрата пропорциональна модулю значения). На рис. б показана матрица W , полученная с помощью байесовского метода главных компонент. Как видно, было выделено три направления, соответствующих загаданным направлениям с наибольшей дисперсией.

PCA + EM = смесь PCA

Рассмотрим следующую вероятностную модель:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k p_k(\mathbf{x} | W_k, \sigma_k^2, \boldsymbol{\mu}_k) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, W_k W_k^T + \sigma_k^2 I), \quad \sum_{k=1}^K \pi_k = 1, \quad \pi_k \geq 0.$$

Эта модель представляет собой смесь нормальных распределений, в которой матрицы ковариации задаются специальным образом. Введем эквивалентную вероятностную модель путем добавления скрытых переменных $z_n \in \{1, \dots, K\}$ для каждого объекта \mathbf{x}_n , отвечающих за номер компоненты смеси:

$$\begin{aligned} p(z_n = k) &= \pi_k, \\ p(\mathbf{x}_n | z_n = k) &= p_k(\mathbf{x}_n). \end{aligned}$$

PCA + EM = смесь PCA

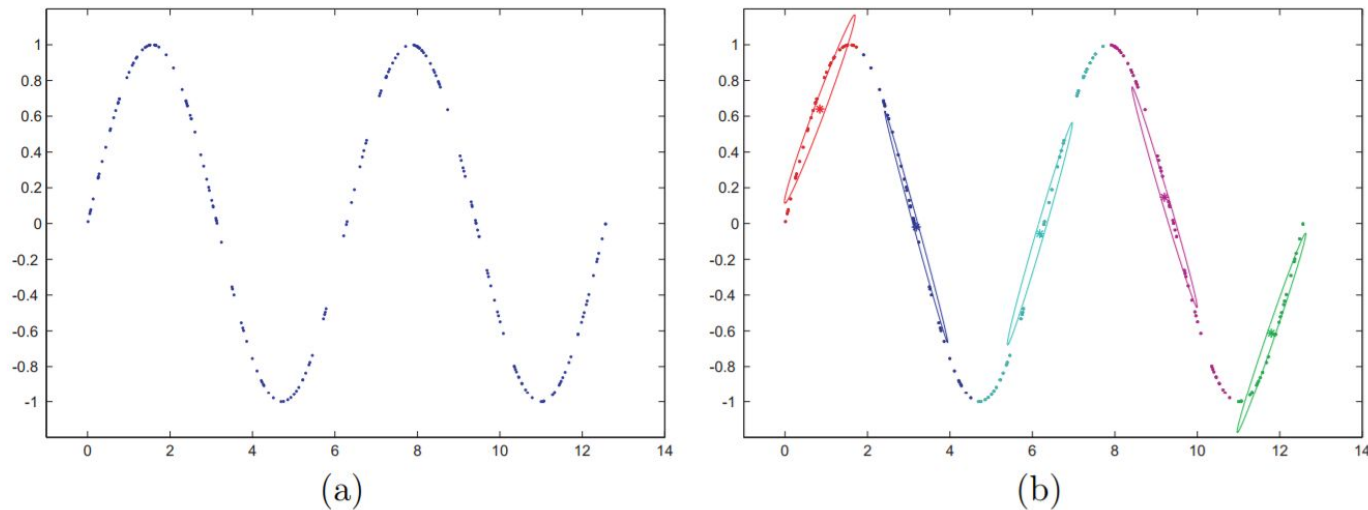


Рис. 6: Кластеризация двухмерной выборки (рис. а) на 5 кластеров с помощью смеси главных компонент (рис. b). Цветами обозначены объекты соответствующих кластеров. Кроме того, показаны центры и эллипсы рассеивания для каждой компоненты смеси.

PCA + EM = VAE

$$p(\mathbf{x}|\mathbf{t}) = \mathcal{N}(\mathbf{x} | W\mathbf{t} + \boldsymbol{\mu}, \sigma^2 I), \quad p(\mathbf{t}) = \mathcal{N}(\mathbf{t} | \mathbf{0}, I).$$

Заменяем на DNN(t)

В следующий раз

- Mean-field approximation
- Variational Bayes
- MCMC
- Непараметрические Байесовские методы
- Прикладные примеры
- ...

Спасибо за внимание!