

**«Машинное обучение и анализ данных»**

# **Байесовский подход к машинному обучению**

**Александр Дьяконов**

**21 апреля 2021 года**

## План

### Формула Байеса

**Оптимальное решение задач классификации - байесовский алгоритм**

случай нормальных распределений

линейный дискриминантный анализ (LDA)

квадратичный дискриминантный анализ (QDA)

Наивный байес (naïve Bayes)

**Байесовский подход в машинном обучении**

Подбрасывание нечестной монетки

Линейная регрессия

MAP

Байесовские оценки параметров: точечные, интервальные

Иерархические модели

**Что такое случайность**

## Задача

**Тест на болезнь «зеленуху» имеет вероятность ошибки 0.1 (как позитивной, так и негативной), зеленухой болеет 10% населения. Какая вероятность того, что человек болен зеленухой, если у него позитивный результат теста?**

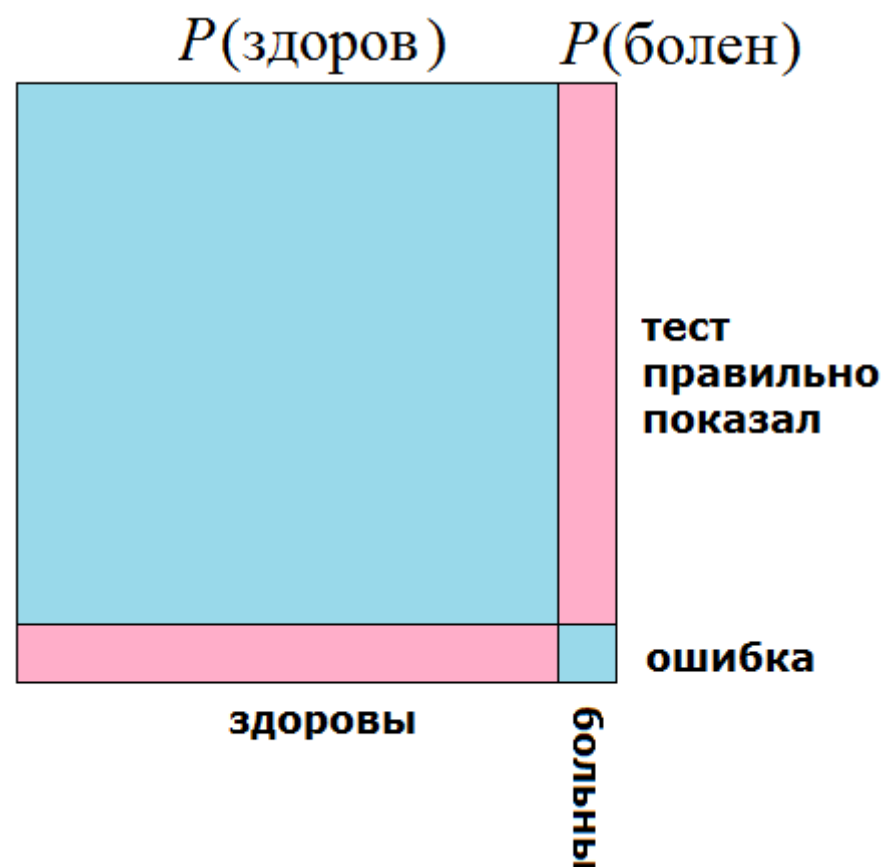
$$P(\text{болен} \mid +) = \frac{P(+ \mid \text{болен})P(\text{болен})}{P(+ \mid \text{болен})P(\text{болен}) + P(+ \mid \text{здоров})P(\text{здоров})}$$

$$\frac{0.9 \cdot 0.1}{0.9 \cdot 0.1 + 0.1 \cdot 0.9} = 0.5$$

**Формула Байеса**

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

## Задача

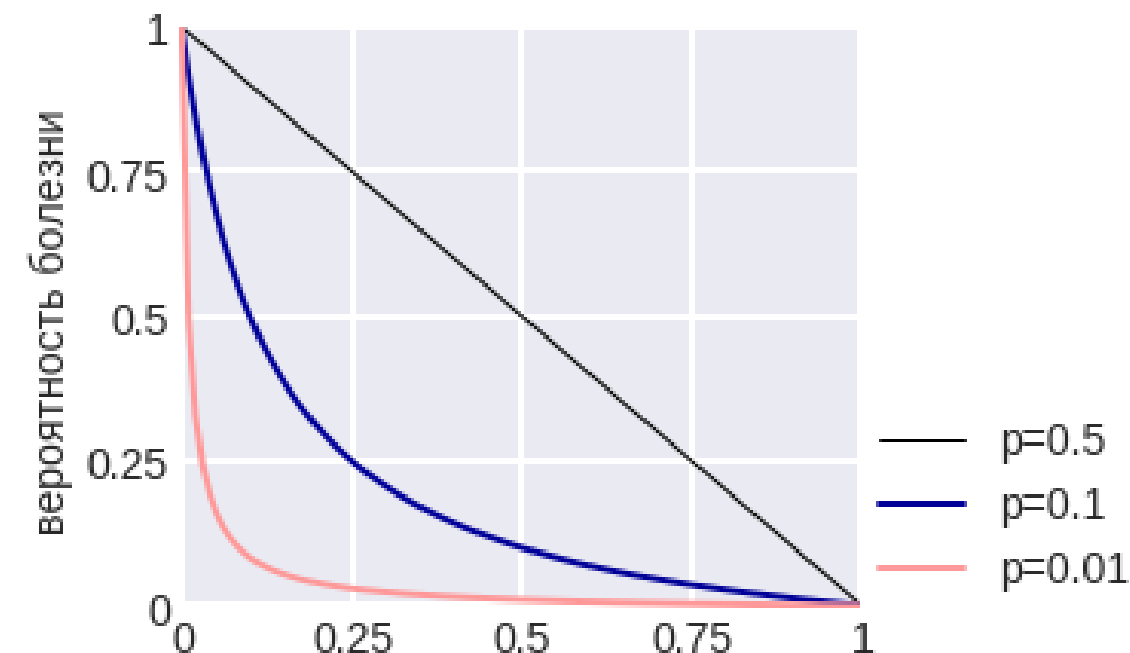
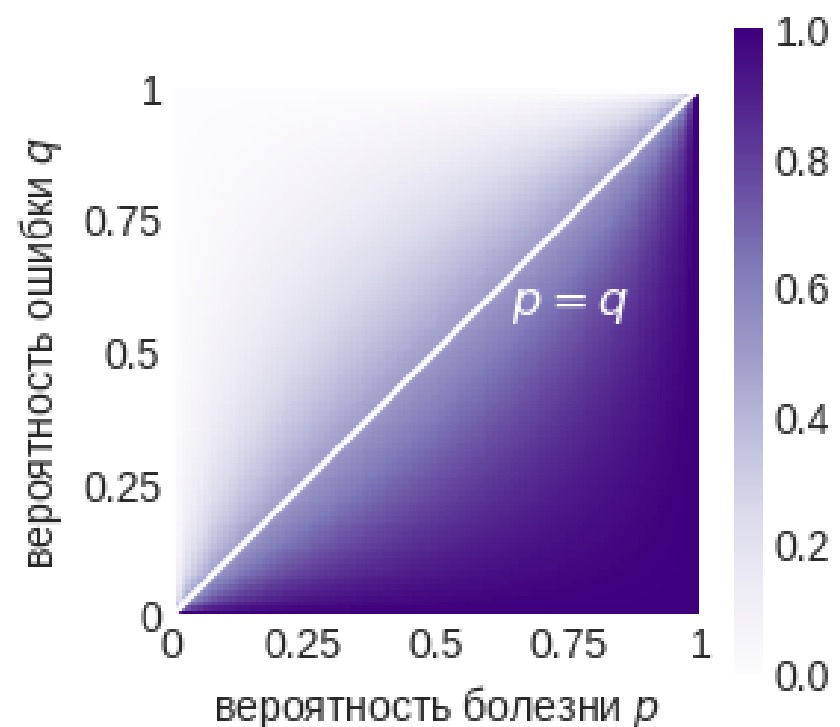


$q$  – вероятность неправильного ответа прибора

$p$  – вероятность заболеть зеленухой

$$P(\text{болен} \mid +) = \frac{(1-q)p}{(1-q)p + q(1-p)} = \frac{p - qp}{p + q - 2qp}$$

## Задача



При  $p = q$  искомая вероятность 0.5

**Чтобы диагностировать очень редкие заболевания нужен сверхточный прибор!!!**

**Нет ли здесь где-то обмана? Почему в жизни мы больше верим тестам?**

## Оптимальное решение задач классификации

Если можем оценить вероятности принадлежности к классам  
(conditional class probabilities)

$$p_k(x) = P(y = k | x), k = 1, 2, \dots, l,$$

тогда **байесовский оптимальный классификатор**

$$a(x) = \arg \max_k (p_k(x))$$

$$p_k(x) = P(y = k | x) = \frac{P(x | y = k) \cdot P(y = k)}{P(x)}$$

**Байесовский оптимальный классификатор имеет наименьшую ошибку в смысле точности классификации**

$P(x | Y = k)$  – плотность распределения (density) объектов класса

$P(Y = k)$  – априорная вероятность (prior probability) класса

## Байесовский оптимальный классификатор

Случай двух классов:



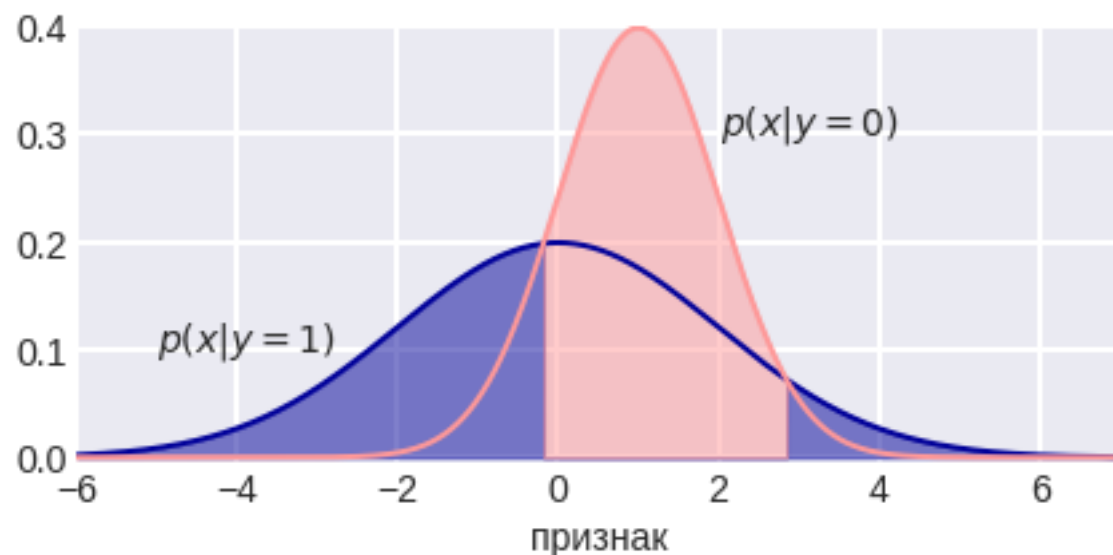
$$p_1(x) = \frac{p(x | y = 1) \cdot P(y = 1)}{P(x)}$$

$$p_0(x) = \frac{p(x | y = 0) \cdot P(y = 0)}{P(x)}$$

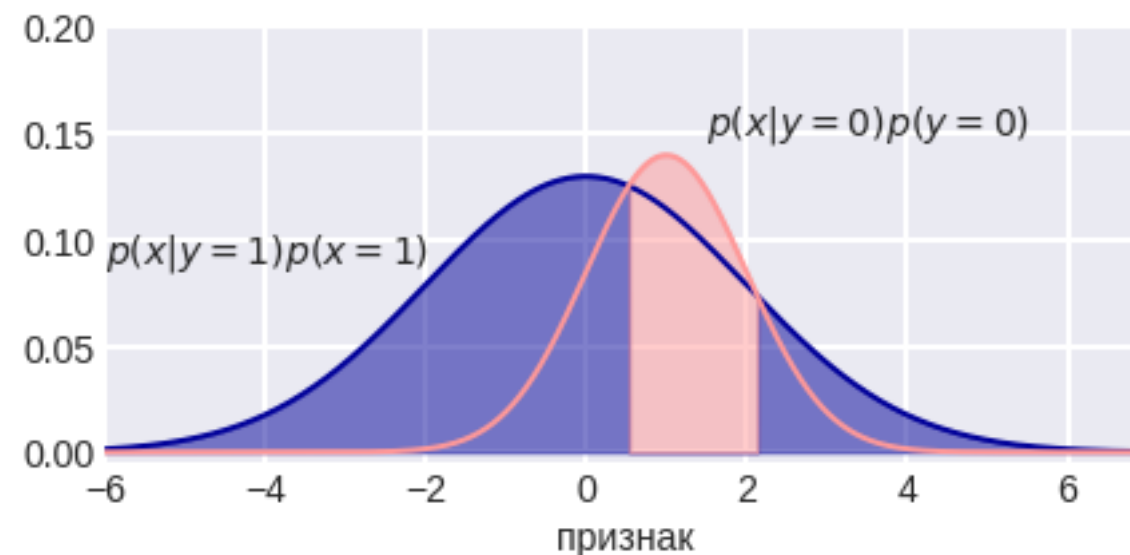
$$p_1(x) <> p_0(x) \Leftrightarrow p(x | y = 1) \cdot P(y = 1) <> p(x | y = 0) \cdot P(y = 0)$$

$P$  – вероятность,  $p$  – плотность

## Частный случай одномерных нормальных распределений



**плотности классов**



$\propto$  **апостериорные вероятности**

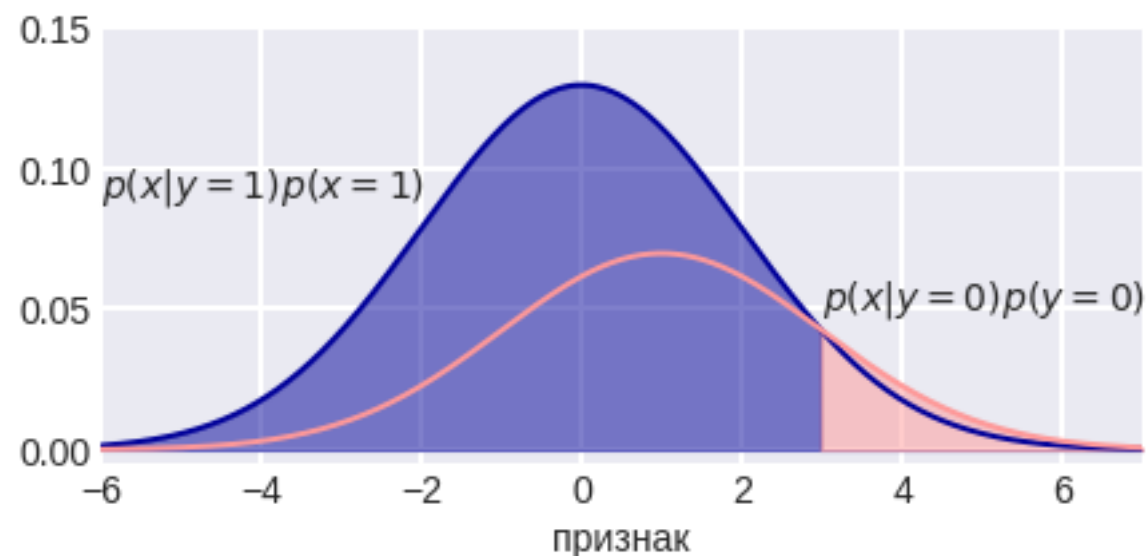
$$p(x | y = k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

$$\frac{p(x | y = 1)}{p(x | y = 0)} \lessgtr \frac{P(y = 0)}{P(y = 1)}$$

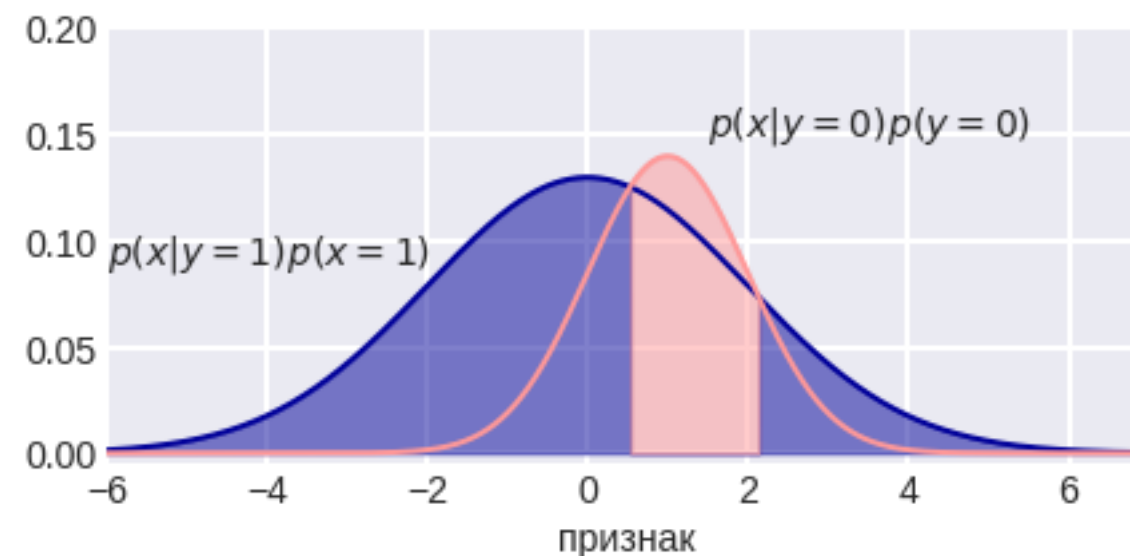
$$p(x | y = 1) \cdot P(y = 1) \lessgtr p(x | y = 0) \cdot P(y = 0) \quad \frac{(x - \mu_0)^2}{\sigma_0^2} - \frac{(x - \mu_1)^2}{\sigma_1^2} \lessgtr \ln \left( \frac{P(y = 0) \cdot \sigma_1}{P(y = 1) \cdot \sigma_0} \right)$$



## Частный случай одномерных нормальных распределений

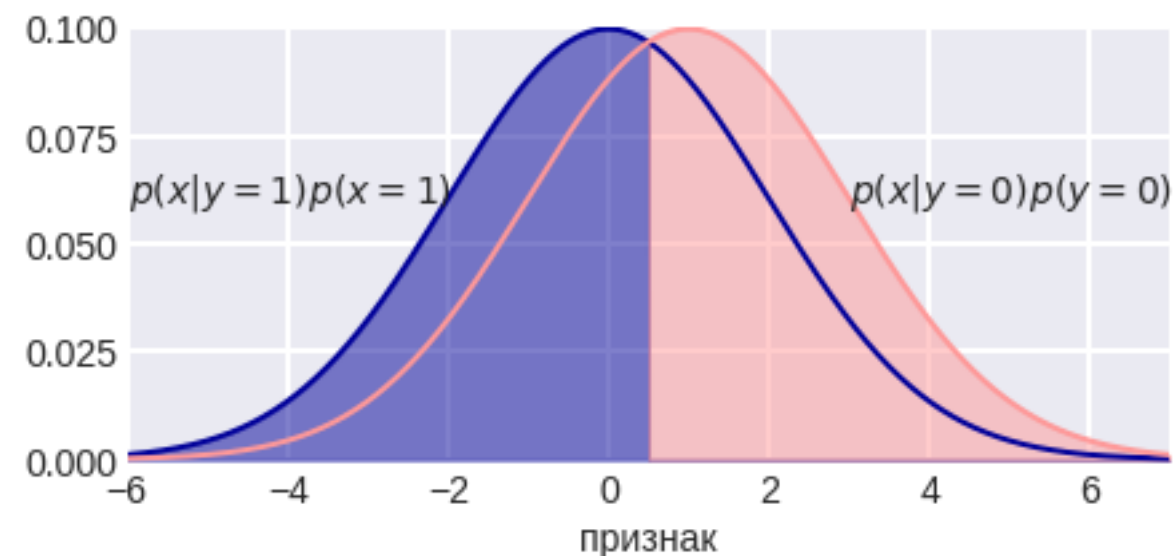


если  $\sigma_0 = \sigma_1$ ,  
то разделяющая поверхность  
(decision boundary) – точка



если  $\sigma_0 \neq \sigma_1$ ,  
то разделяющая поверхность –  
две точки

## Частный случай одномерных нормальных распределений



если  $\sigma_0 = \sigma_1$  и классы равновероятны,

то разделяющая поверхность  $\frac{\mu_1 + \mu_2}{2}$

## Оценка параметров

**Как оценить описанные выше параметры?**

$$P(y = k) = \frac{|\{i \mid y_i = k\}|}{m} \text{ – процент объектов класса } k \text{ из известных}$$

$$\mu_k = \frac{1}{|\{i \mid y_i = k\}|} \sum_{i: y_i = k} x_i$$

$$\Sigma_k = \frac{1}{|\{i \mid y_i = k\}| - 1} \sum_{i: y_i = k} (x_i - \mu_k)(x_i - \mu_k)^T$$

## Частный случай многомерных нормальных распределений

$$p(x \mid y = k) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

$$\frac{p(x \mid y = 1)}{p(x \mid y = 0)} \lessgtr \frac{P(y = 0)}{P(y = 1)}$$

$$\frac{\frac{1}{(2\pi)^{p/2} |\Sigma_1|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)}}{\frac{1}{(2\pi)^{p/2} |\Sigma_0|^{1/2}} e^{-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)}} \lessgtr \frac{P(y = 0)}{P(y = 1)}$$

## Частный случай многомерных нормальных распределений

$$\frac{e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)}}{e^{-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)}} \langle \rangle \frac{|\Sigma_1|^{1/2} P(y=0)}{|\Sigma_0|^{1/2} P(y=1)}$$

$$+(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0) - (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) \langle \rangle 2 \ln \frac{|\Sigma_1|^{1/2} P(y=0)}{|\Sigma_0|^{1/2} P(y=1)}$$

$$\frac{1}{2} x^T (\Sigma_0^{-1} - \Sigma_1^{-1}) x - x^T (\Sigma_0^{-1} \mu_0 - \Sigma_1^{-1} \mu_1) \langle \rangle \mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0 + 2 \ln \frac{|\Sigma_1|^{1/2} P(y=0)}{|\Sigma_0|^{1/2} P(y=1)}$$

## Частный случай – линейный дискриминантный анализ (LDA)

$$\frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x - x^T(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) \lessgtr \text{const}$$

**если ковариационные матрицы равны**

$$x^T w \lessgtr \text{const}$$

**– линейная разделяющая поверхность!**

так, кстати, выводится логистическая регрессия (**была иллюстрация**)

приём в ML – часто ковариационные матрицы полагают равными,  
чтобы оценивать меньше параметров (parameter sharing)

**Логистическая регрессия более робастна,  
лучше работает, когда априорные предположения не выполняются**

## LDA: логистическая регрессия

маленькое отступление...

**Решаем бинарную задачу классификации**  
**Предполагаем, что**

$$p(y = 1 | x) = \frac{1}{1 + \exp(-w^T x)} = \frac{\exp(+w^T x)}{1 + \exp(+w^T x)}$$
$$p(y = 0 | x) = 1 - p(y = 1 | x) = \frac{1}{1 + \exp(+w^T x)}$$
$$\frac{p(y = 1 | x)}{p(y = 0 | x)} = \exp(-w^T x)$$

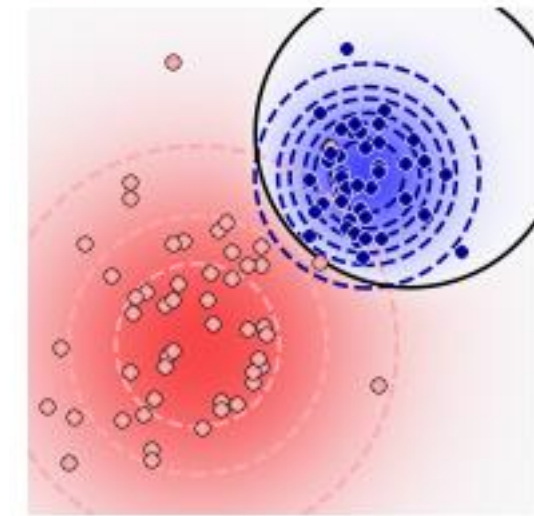
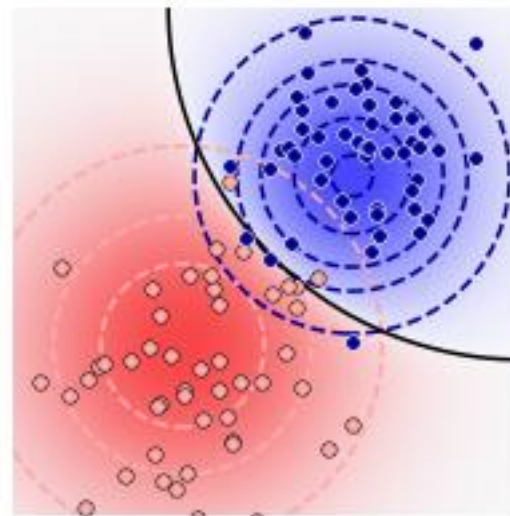
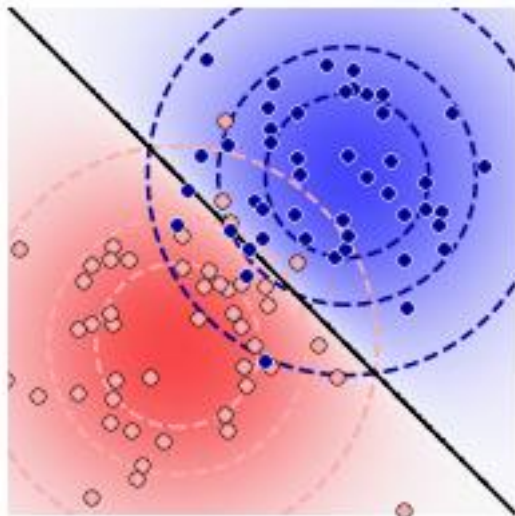
**т.е. формально «чёткая классификация» – линейная модель**

## Частный случай – квадратичный дискриминантный анализ (QDA)

$$\frac{1}{2}x^T W x - x^T w <> \text{const}$$

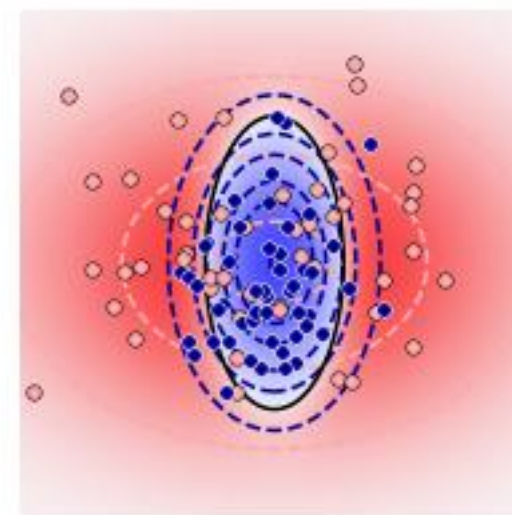
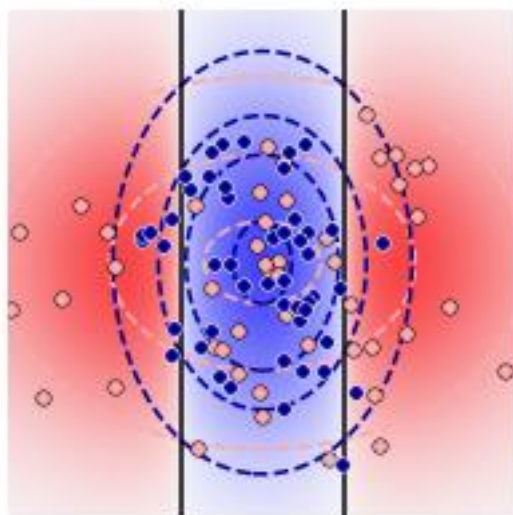
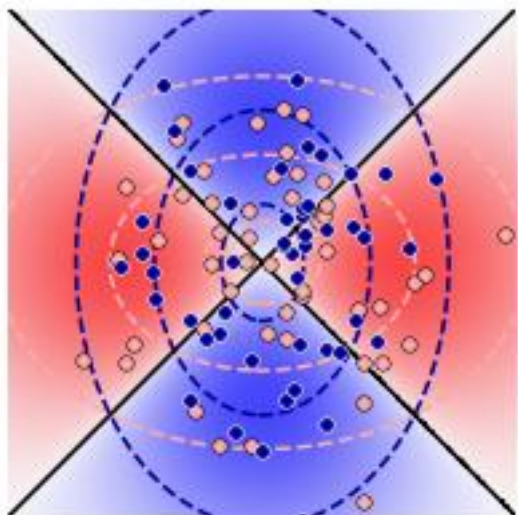
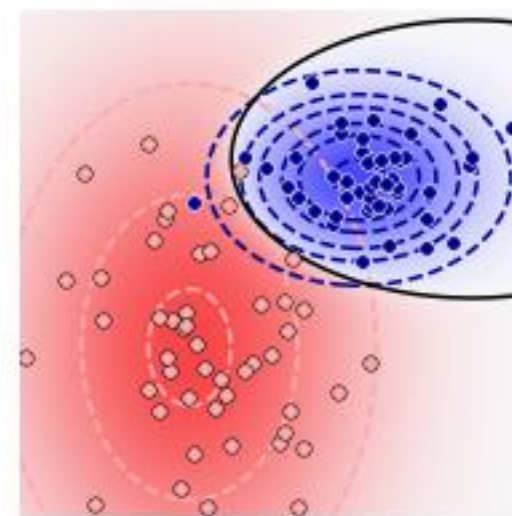
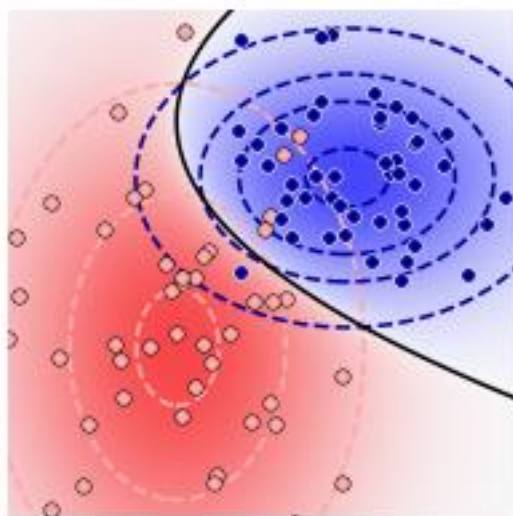
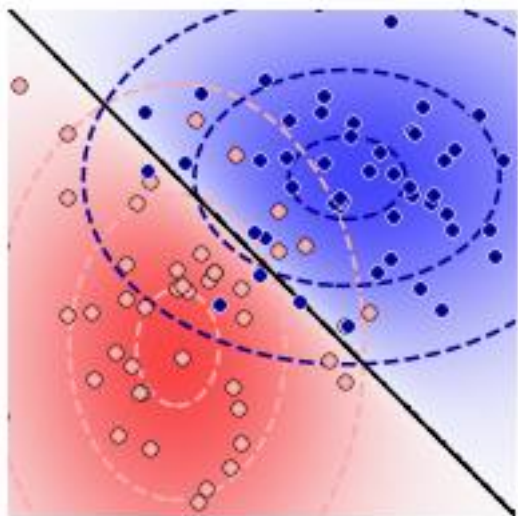
**не просто разделяемая поверхность, а теоретически обоснованная!**

### Gaussian Bayes Classifier





Gaussian Bayes Classifier



## Минимизация среднего риска

$$\sum_k \sum_s \lambda_{ks} P(y = k) P(a(x) = s \mid y(x) = k)$$

$\lambda_{ks}$  – штраф за отнесение алгоритмом объекта  $k$ -го класса к  $s$ -му  
 $P(y = k) P(a(x) = s \mid y(x) = k)$  – вероятность такого события

### Оптимальный классификатор:

$$\begin{aligned} s &= \arg \min_s \sum_k \lambda_{ks} \cdot P(y = k \mid x) = \\ &= \arg \min_s \sum_k \lambda_{ks} \cdot P(y = k) \cdot P(x \mid y = k) \end{aligned}$$

## Наивный байес (naïve Bayes)

$$p(x = (X_1, \dots, X_n) | y) = \prod_{j=1}^n p_j(X_j | y)$$

**Плотность = произведение плотностей по всем признакам**  
(т.е. предполагается независимость признаков)

**Для нормальных распределений ~ диагональность матриц ковариаций**

**Сильное предположение, но часто так получается неплохое решение!**  
(~регуляризация – поиск решений среди простых)

**Бонус: оценивать не многомерную, а одномерные плотности**  
(меньше требования к объёму выборки)

считается, что наивный Байес неплох для больших размерностей...  
пример применения наивного Байеса – mean target encoding

## Байесовский алгоритм – итог

- + теоретически наилучший алгоритм**
- нужна вероятностная природа данных, знание распределений, их точная оценка**  
**тут, в некотором смысле, накопление ошибок**
- + большинство естественных решений – на байесовском подходе**
- + получили теоретические обоснования простых алгоритмов**  
**(линейных и квадратичных – сложнее не надо...)**
- + возможности для разделения параметров, «наивизации»**

## Байесовский подход в машинном обучении

$$p(\theta | D) = \frac{p(D | \theta) \cdot p(\theta)}{p(D)} = \frac{p(D | \theta) \cdot p(\theta)}{\int_{\theta} p(D | \theta) p(\theta) d\theta}$$

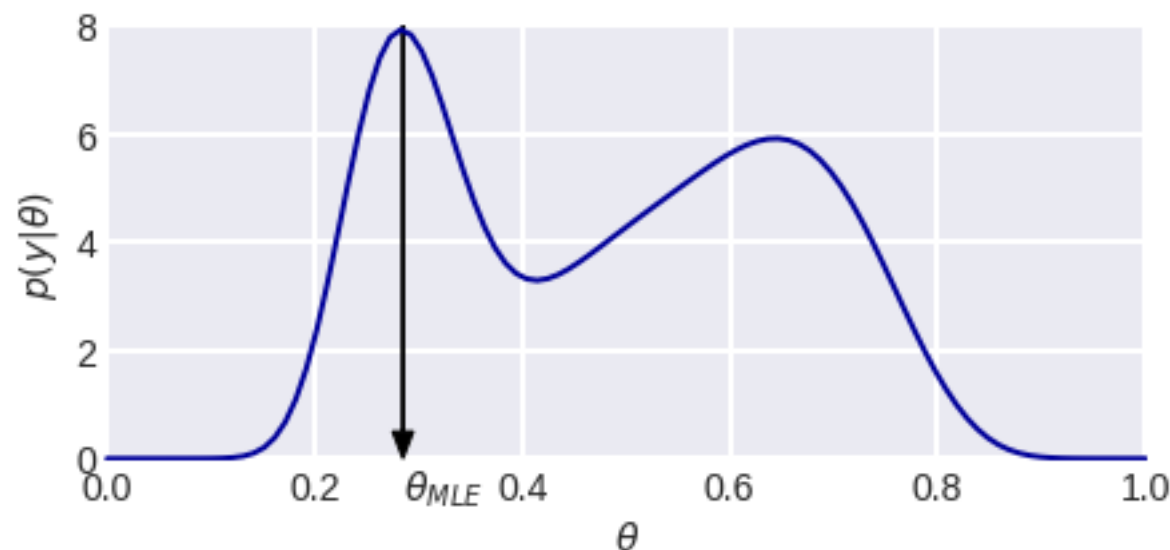
- $D$  – данные
- $p(D)$  – вероятность данных (Marginal likelihood/ Model evidence)  
нормировка, чтобы получалась плотность
- $\theta$  – параметр (модели)
- $p(\theta)$  – априорная вероятность/распределение (prior)
- $p(D | \theta)$  – правдоподобие (likelihood)
- $p(\theta | D)$  – апостериорная вероятность/распределение (posterior)

## Метод максимального правдоподобия

**Независимые одинаково распределённые с.в.:**  $y_1, \dots, y_m$

**Распределение известно с точностью до параметра:**  $p(y | \theta)$   
(можно пока считать, что это плотность)

**Правдоподобие (likelihood):**  $p(y | \theta) = p(y_1, \dots, y_m | \theta) = \prod_{i=1}^m p(y_i | \theta) \rightarrow \max$



**Находим значение параметра(ов),  
которое делает наблюдаемые данные  
максимально правдоподобными**

**MLE = Maximum Likelihood Estimation**

**Пример: подбрасывание нечестной монетки (coin-toss problem)**

$$y_i = \begin{cases} 1, & \text{орёл,} \\ 0, & \text{решка.} \end{cases}$$

**Делаем предположение, что подчиняется  
распределению Бернулли:**

$$p(y_i | \theta) = \theta^{y_i} (1 - \theta)^{1 - y_i}$$

**Логарифм правдоподобия:** 
$$\sum_{i=1}^m \log p(y_i | \theta) = \sum_{i=1}^m y_i \log \theta + (1 - y_i) \log(1 - \theta)$$

**на что похоже выражение под суммой?**

**Если взять производную и приравнять к нулю...**

$$\hat{\theta}_{\text{MLE}} = \frac{1}{m} \sum_{i=1}^m y_i$$

**~ оценка максимального правдоподобия вероятности выпадения орла ч.т.д.?**

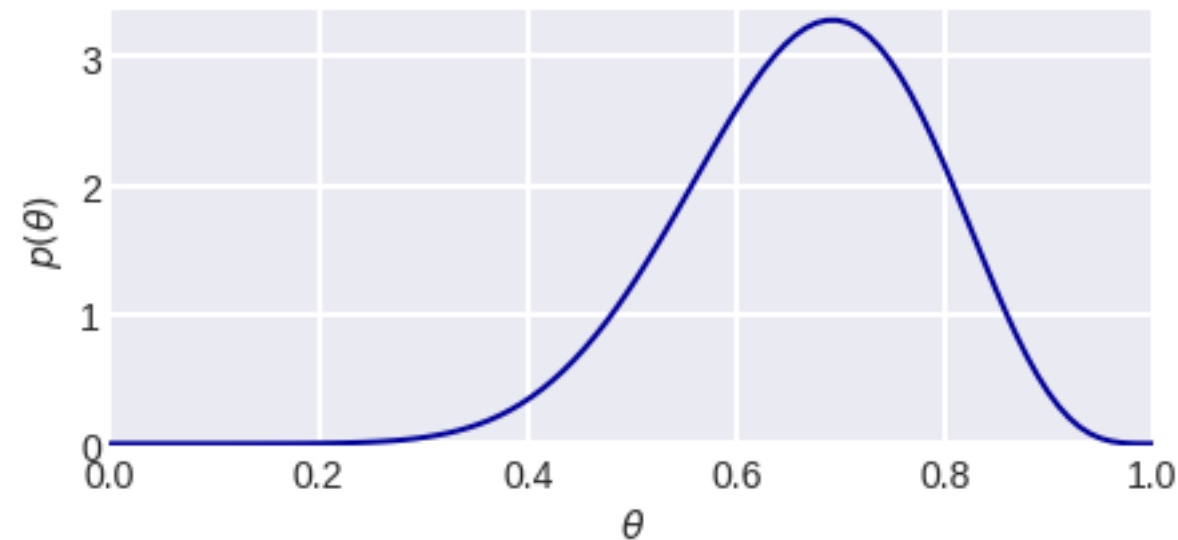
## Что не так с оценкой MLE

- **переобучение...**
- **ненадёжность при малом числе экспериментов**
- **нет возможности внести априорные предположения о вероятности**



## Байесовский подход в нашем примере

Пусть есть априорное распределение параметра  $p(\theta)$  (prior):



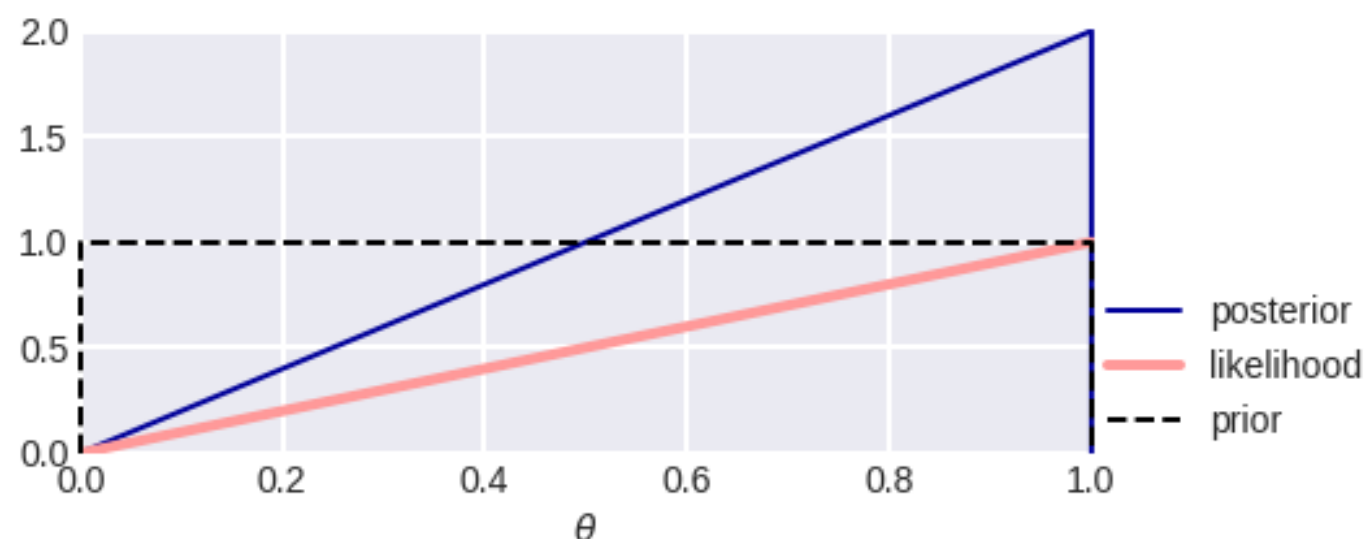
**это будет формализация априорных знаний + регуляризация!**

**Замечание:** равномерное распределение – отсутствие априорных знаний...

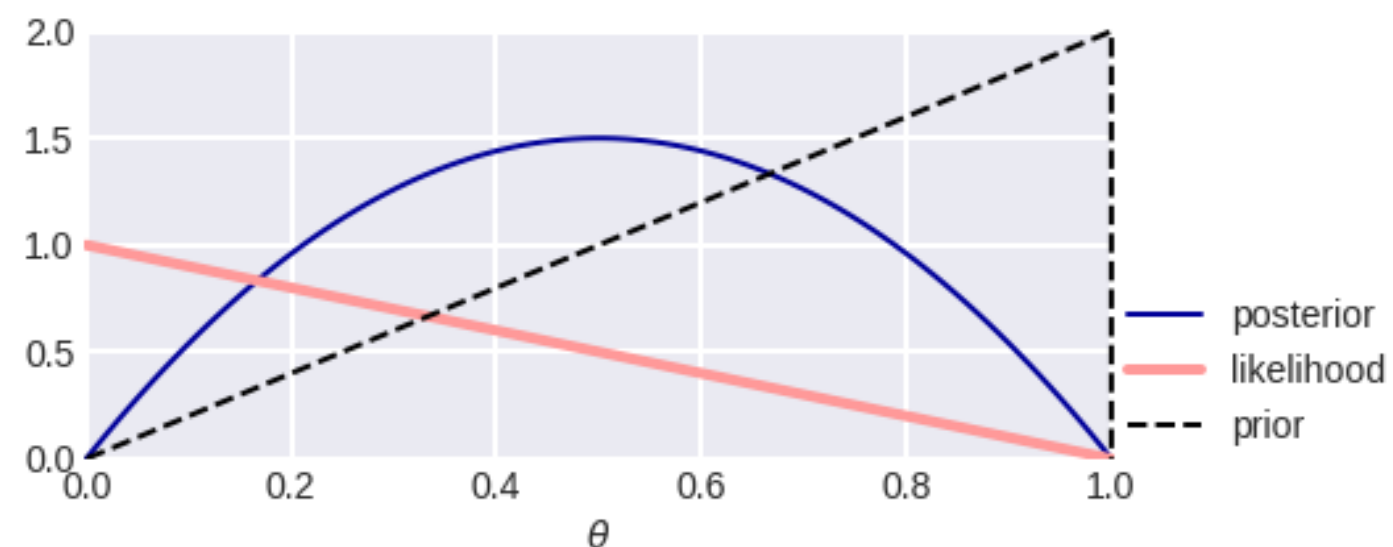
## Байесовский подход в нашем примере

## Формула Байеса (Bayes rule)

$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)}$$

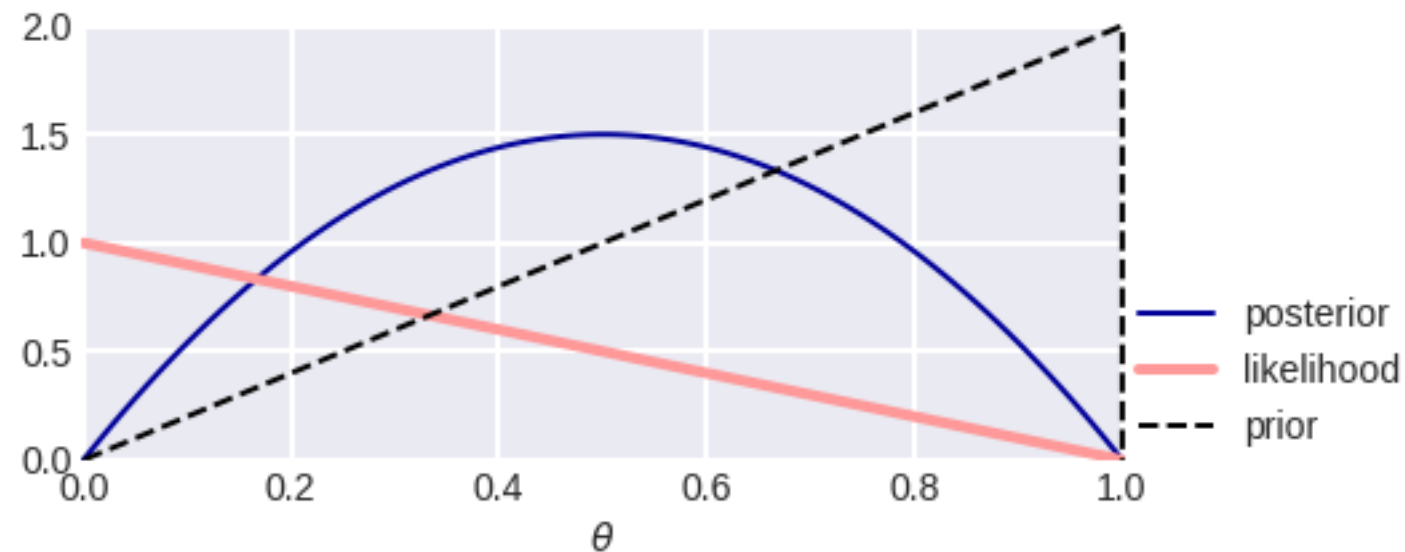


ничего не знаем ( $p(\theta) = 1$ )  $\rightarrow$   
один орёл ( $p(y_1 = 1 | \theta) = \theta^1 (1 - \theta)^{1-1} = \theta$ )  $\rightarrow$   
 $p(\theta | y_1 = 1) = 2\theta$



знания как будто был один орёл (или был в  
1м эксперименте)  $p(\theta | y_1 = 1) = 2\theta \rightarrow$   
одна решка  
( $p(y_2 = 0 | \theta) = \theta^0 (1 - \theta)^{1-0} = 1 - \theta$ )  $\rightarrow$   
 $p(\theta | y_2 = 0) = 6\theta(1 - \theta)$

## Байесовский подход в нашем примере



$$p(\theta \mid y_1 = 1, y_2 = 0) = 6\theta(1 - \theta)$$

**В итоге получим распределение на множестве значений параметра!**

**Как выбрать конкретное значение параметра – дальше – пока моду...**

## Байесовский подход в нашем примере

$$\text{MLE: } p(y | \theta) \rightarrow \max$$

$$\text{MAP: } p(\theta | y) \rightarrow \max$$

Теперь максимизируем апостериорную вероятность (posterior distribution)

### MAP (Maximum A Posteriori)

$$\log p(\theta | y) = \log p(y | \theta) + \log p(\theta) - \log p(y)$$

$$\log p(y | \theta) + \log p(\theta) \rightarrow \max$$

второе слагаемое – регуляризатор,

т.е. это как бы MLE + регуляризация

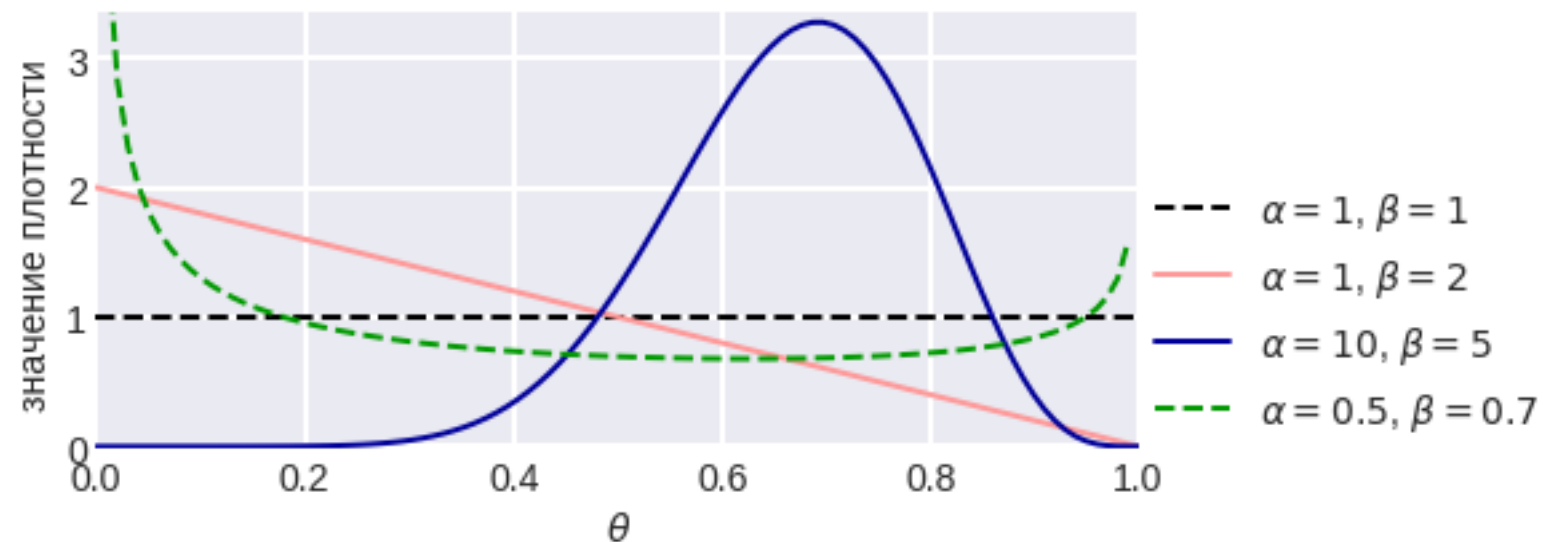
При равномерном распределении регуляризатор исчезает

## Байесовский подход в примере с монетой

Априорное Бета-распределение  $\theta \sim \text{Beta}(\alpha, \beta)$ :

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

$\alpha, \beta$  – гиперпараметры априорного распределения



**Байесовский подход в примере с монетой**

$$\begin{aligned} & \log p(y | \theta) + \log p(\theta) = \\ &= \sum_{i=1}^m y_i \log \theta + (1 - y_i) \log(1 - \theta) + (\alpha - 1) \log \theta + (\beta - 1) \log(1 - \theta) + \text{const} \end{aligned}$$

пренебрегаем константами

**Берём производную и приравниваем к нулю:**

$$\hat{\theta}_{\text{MAP}} = \frac{\sum_{i=1}^m y_i + \alpha - 1}{m + \alpha + \beta - 2}$$

узнаём «сглаживание по Лапласу»

**Для равномерного априорного распределения ( $\alpha = \beta = 1$ ) оценки MLE и MAP совпадают**

**Оценку можно трактовать в терминах наличия априорных экспериментов**

## Особенности байесовского подхода

**ММП даёт точечную оценку параметра (конкретное значение).**

**Байесовский подход – целое распределение параметра – апостериорное!**

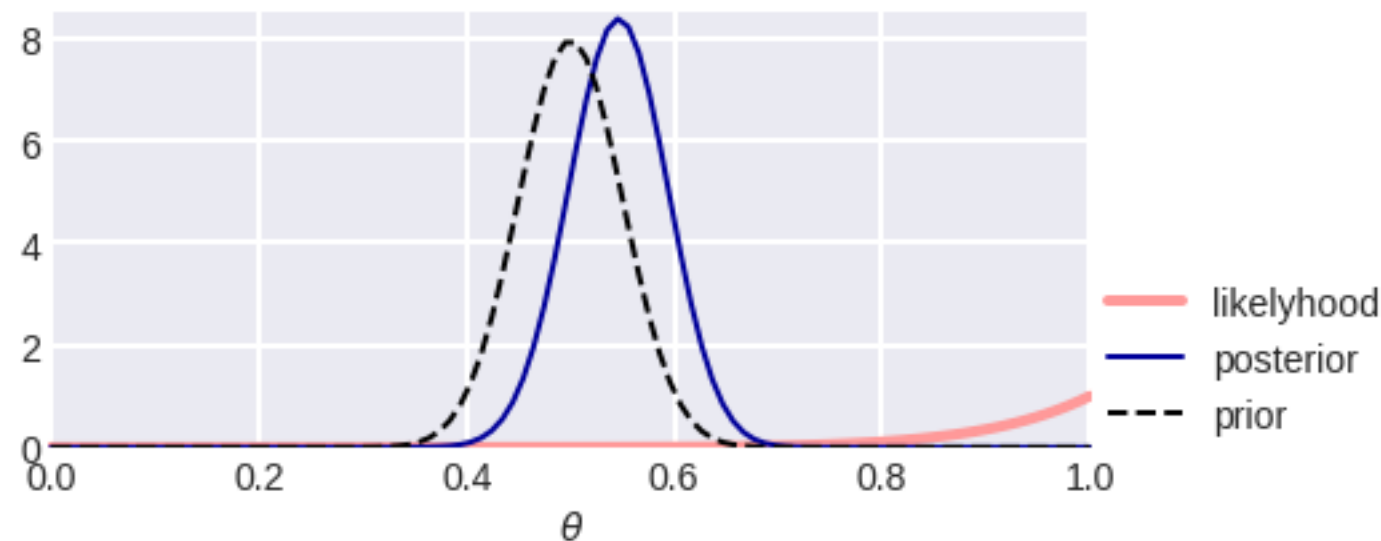
$$p(\theta | y) = \frac{p(y | \theta) p(\theta)}{p(y)}$$

**априорное распределение → апостериорное распределение**

**Красиво и эффективно БП применяется,  
если апостериорное распределение из того же семейства, что и априорное  
(и удаётся найти формулу для пересчёта параметров)**

## Особенности байесовского подхода

**Может применяться в онлайн-режиме (данные поступают батчами)**



**верили в честность монетки → 10 орлов подряд → изменили веру**

$$p(\theta) = \text{Beta}(50, 50) \rightarrow p(y | \theta) = \theta^{10} \rightarrow p(\theta | y) = \text{Beta}(60, 50)$$

$$p(\theta_{\text{new}}) = p(\theta | y)$$

**Апостериорное распределение становится априорным для нового батча**



## Особенности байесовского подхода

Теперь ясно почему связались с бета-распределением:

$$\begin{aligned} p(y | \theta) p(\theta) &= \theta^k (1 - \theta)^{n-k} \cdot C \cdot \theta^{\alpha-1} (1 - \theta)^{\beta-1} = \\ &= C \cdot \theta^{k+\alpha-1} (1 - \theta)^{n-k+\beta-1} \end{aligned}$$

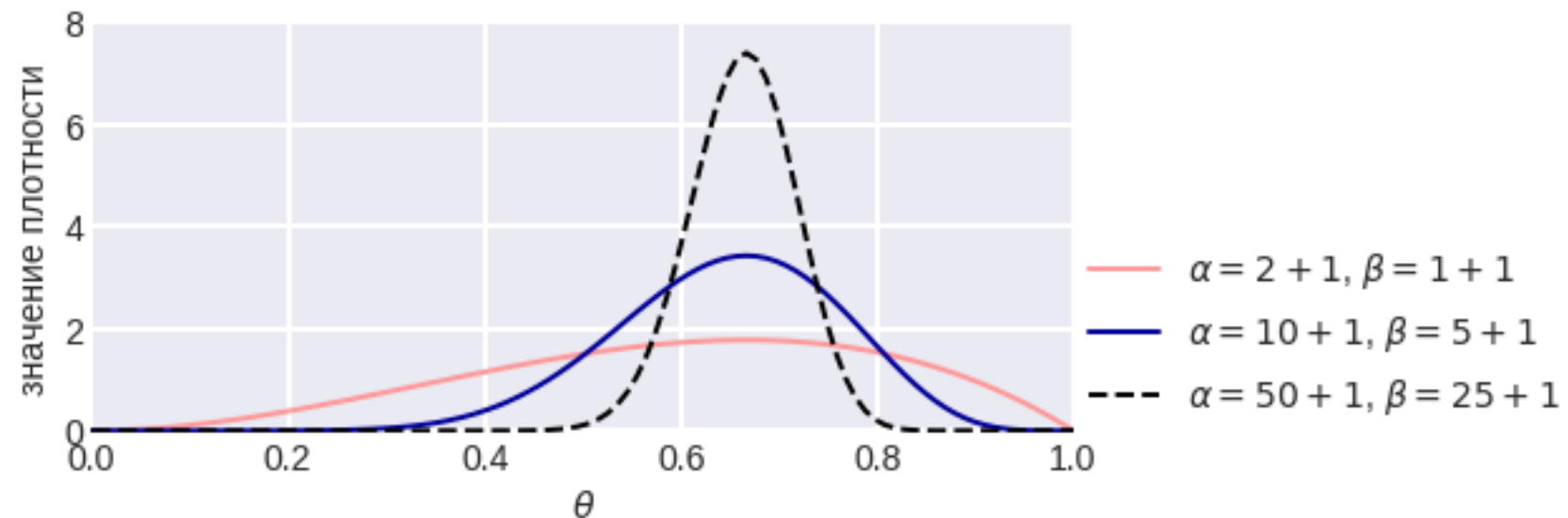
Если это распределение умножить на правдоподобие и нормировать,  
то получим опять это распределение (с другими параметрами)

Такие распределения **сопряжённые**

Бета-распределение сопряжённое к распределению Бернулли  
И интеграл от их произведения (условие нормировки) берётся аналитически

## Особенности байесовского подхода

**С помощью бета-распределения удобно формализовать  
«веру в определённую вероятность»**



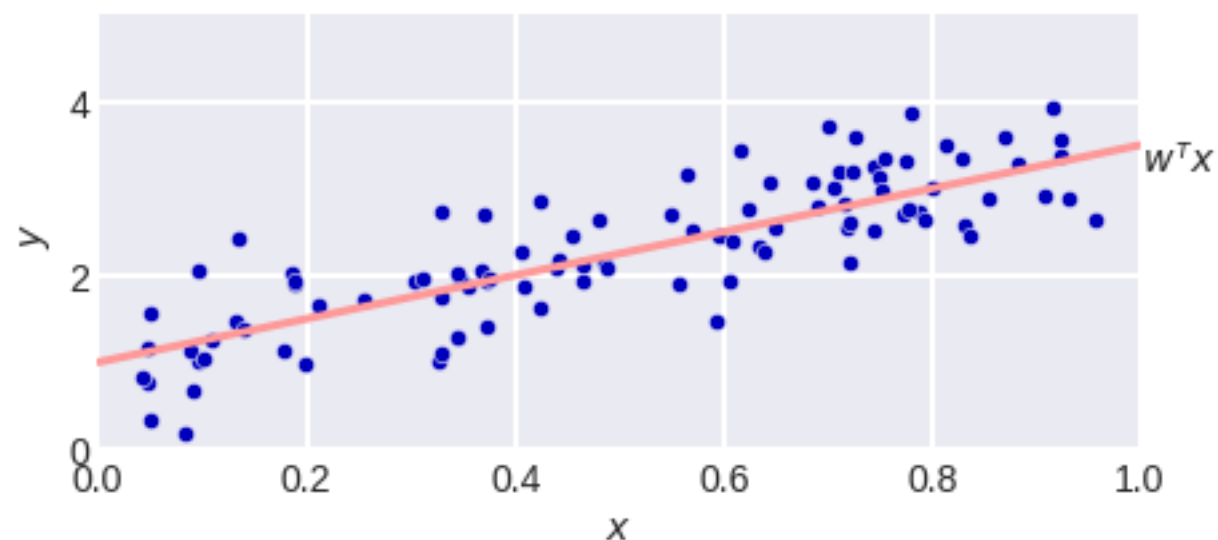
## MLE для линейной регрессии (напоминание)

Пусть

$$y = w^T x + \varepsilon$$

$$\varepsilon \sim \text{norm}(0, \sigma^2)$$

таким образом  $y \sim \text{norm}(w^T x, \sigma^2)$



**MLE для линейной регрессии (напоминание)**

$$p(y | x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y - w^T x)^2}{2\sigma^2}\right]$$

$$\log L(w) = \log \prod_{i=1}^m p(y_i | x_i, w) = \sum_{i=1}^m \left[ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - w^T x_i)^2}{2\sigma^2} \right]$$

**Максимизация правдоподобия (MLE):**

$$\frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - w^T x_i)^2 \rightarrow \min,$$

**т.е. это МНК!**

## Байесовская теория для линейной регрессии

**А теперь введём априорное распределение на веса**

$$p(w) = \text{norm}(w \mid 0, \lambda^{-1}I) = \frac{\lambda^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{\lambda}{2} w^T w\right]$$

**Как всегда логика: «меньше – лучше»**

$$\log p(w \mid D) = \log p(w) + \log p(D \mid w) - \log p(D)$$

**MAP**

$$\begin{aligned} \log p(w \mid D) &\sim \log p(w) + \log p(D \mid w) = \\ &= \log \left[ \frac{\lambda^{n/2}}{(2\pi)^{n/2}} \right] - \frac{\lambda}{2} w^T w + \sum_{i=1}^m \left[ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - w^T x_i)^2}{2\sigma^2} \right] \end{aligned}$$

## Байесовская теория для линейной регрессии

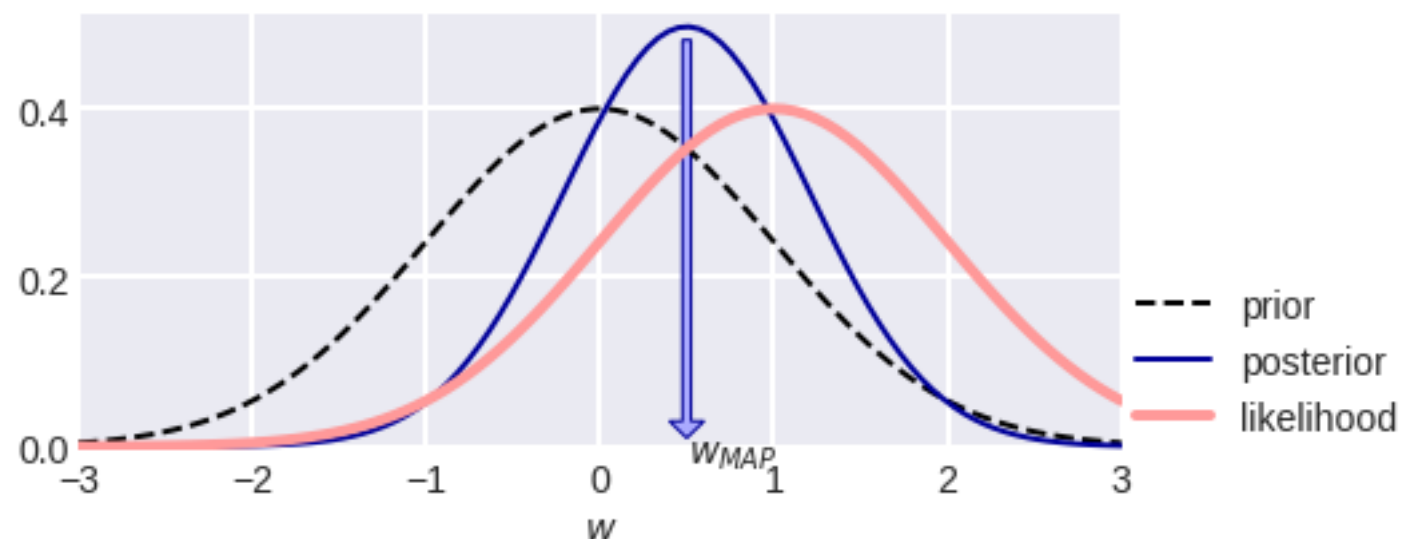
### MAP

$$\frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - w^T x_i)^2 + \frac{\lambda}{2} w^T w \rightarrow \min$$

**т.е. это гребневая регрессия!**

**мы наконец-то (ещё раз!) теоретически обосновали Ridge()!**

**И тут ясен смысл регуляризационного множителя  $\lambda\sigma^2$**



## Байесовская теория для линейной регрессии

- **MLE** – нерегуляризованное решение
- **MAP** – регуляризованное решение

**регуляризация ~ априорное распределение параметров**

**Априорное нормальное распределение сводится к гребневой регрессии  
( $l_2$ -регуляризации)**

**Если использовать априорное распределение Лапласа,  
то получим  $l_1$ -регуляризацию**

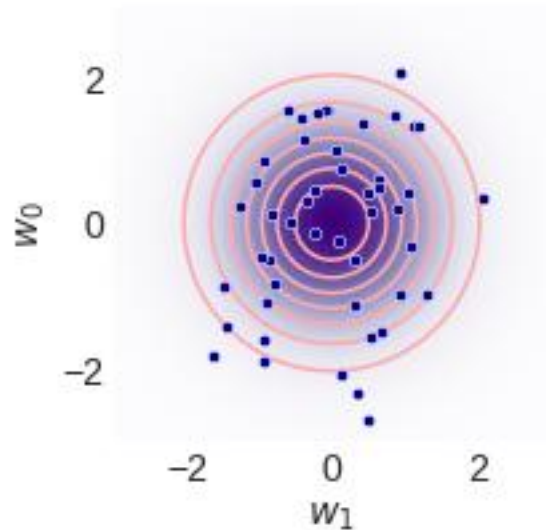
**Аналогично с логистической регрессией!**

## Байесовская теория для линейной регрессии

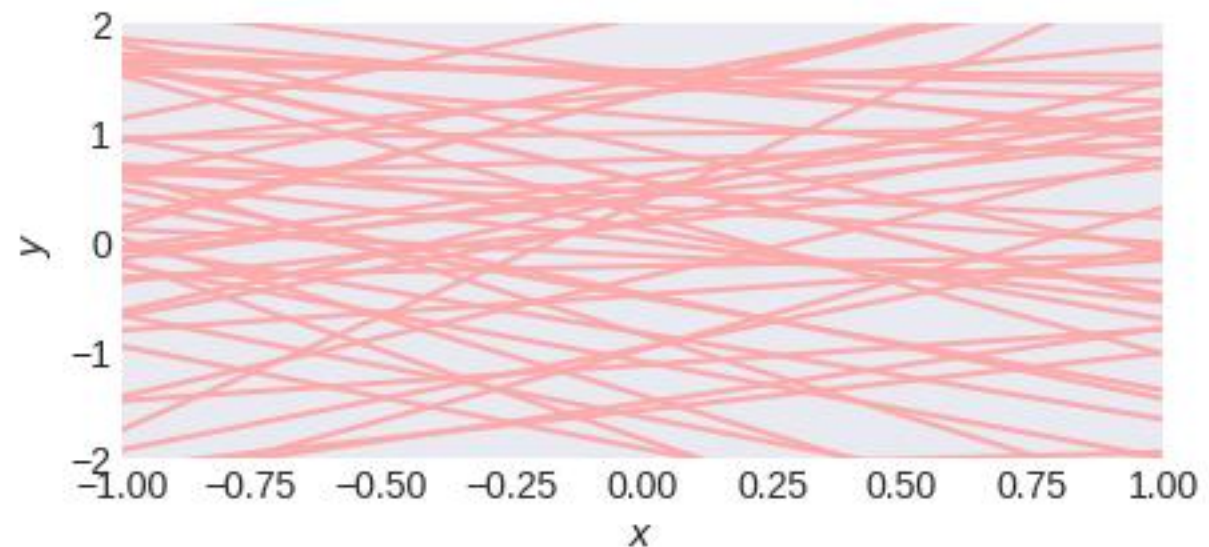
### априорное распределение

$$p(w) = \text{norm}(w | 0, \Sigma_0) = \frac{1}{(2\pi)^{n/2} |\Sigma_0|^{1/2}} \exp\left[-\frac{1}{2} w^T \Sigma_0^{-1} w\right]$$

### распределение на значениях коэффициентов



### пространство данных и сэмплированные модели





## Байесовская теория для линейной регрессии

### MAP

$$\begin{aligned}
 \log p(w | D) &\sim \log p(w) + \log p(D | w) = \\
 &= \text{const} - \frac{1}{2} w^T \Sigma_0^{-1} w - \frac{1}{2\sigma^2} \|Xw - y\|^2 = \\
 &= \text{const} - \frac{1}{2\sigma^2} (w^T X^T X w - 2y^T X w + y^T y) - \frac{1}{2} w^T \Sigma_0^{-1} w = \\
 &= -\frac{1}{2} w^T \left( \frac{1}{\sigma^2} X^T X + \Sigma_0^{-1} \right) w + \frac{2}{2\sigma^2} y^T X w + \text{const} = -\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) + \text{const}
 \end{aligned}$$

если выделить полный квадрат... можно раскрыть последнюю строчку и приравнять

### постериорное распределение

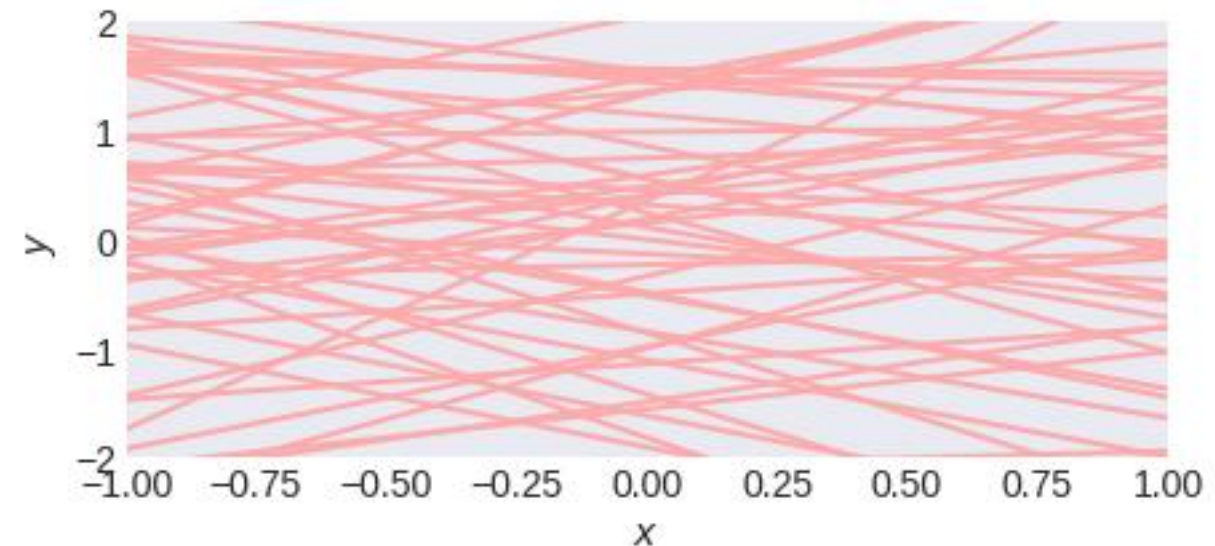
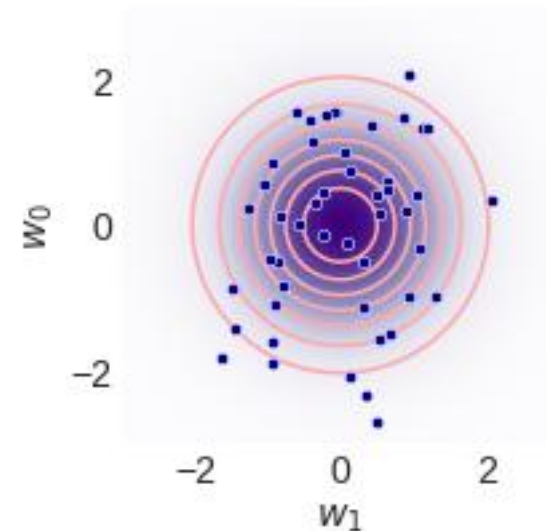
$$P(w | D) = \text{norm}(\sigma^{-2} (\sigma^{-2} X^T X + \Sigma_0^{-1})^{-1} X^T y, (\sigma^{-2} X^T X + \Sigma_0^{-1})^{-1})$$

posterior mean, posterior variance

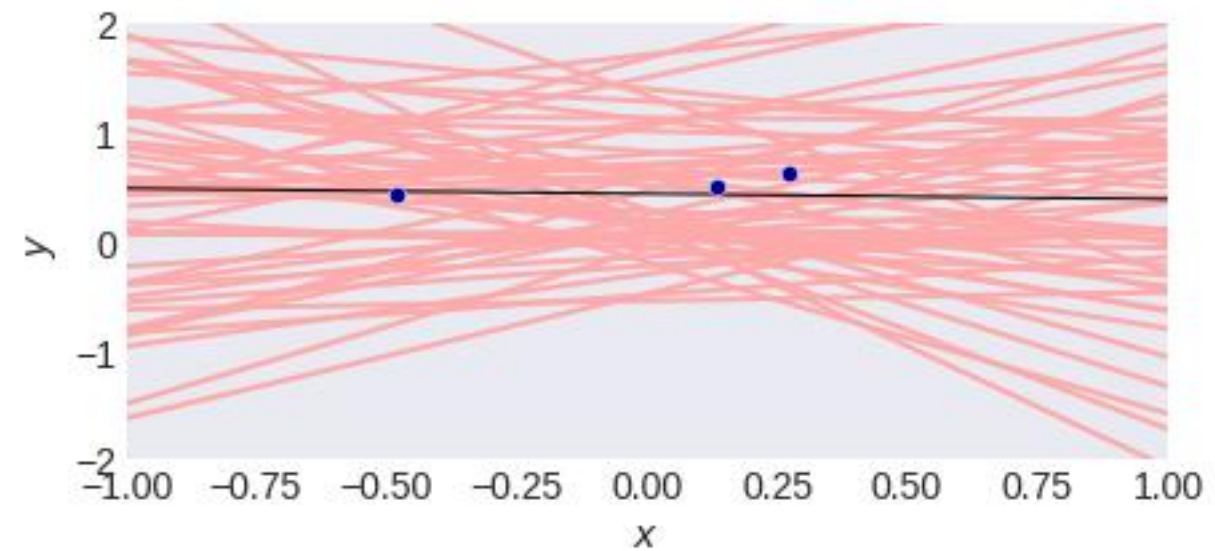
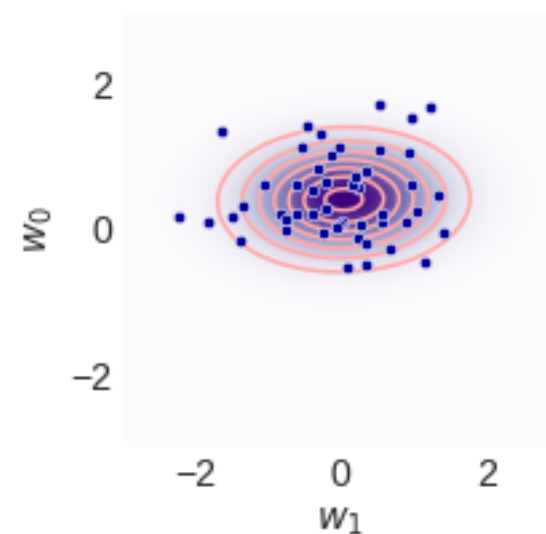
## Байесовская теория для линейной регрессии

$$m = 0$$

только априорное  
нормальное  
распределение  
значений  
коэффициентов

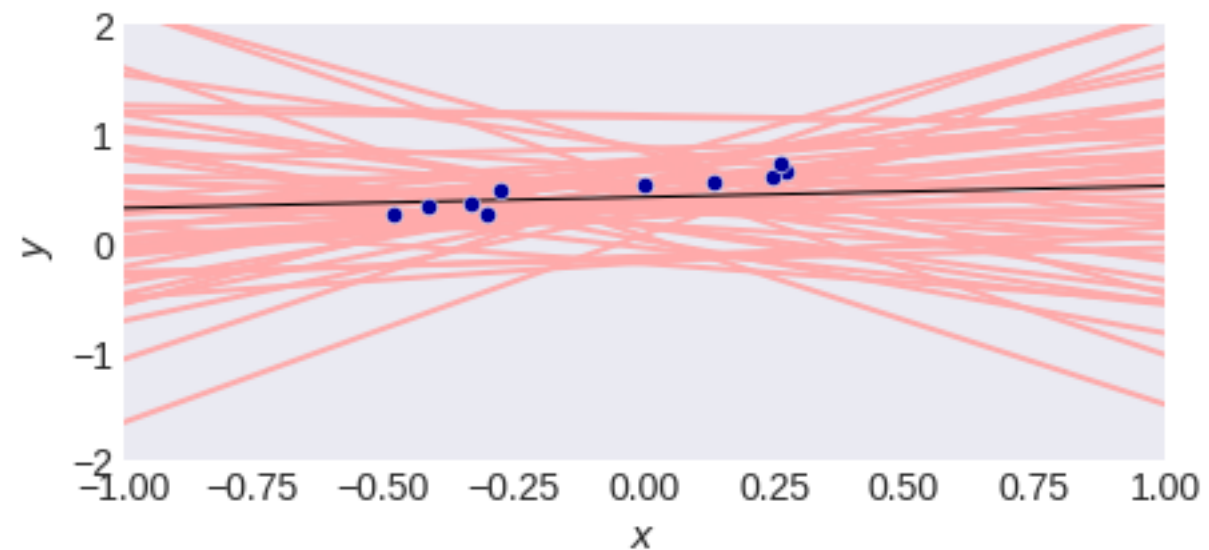
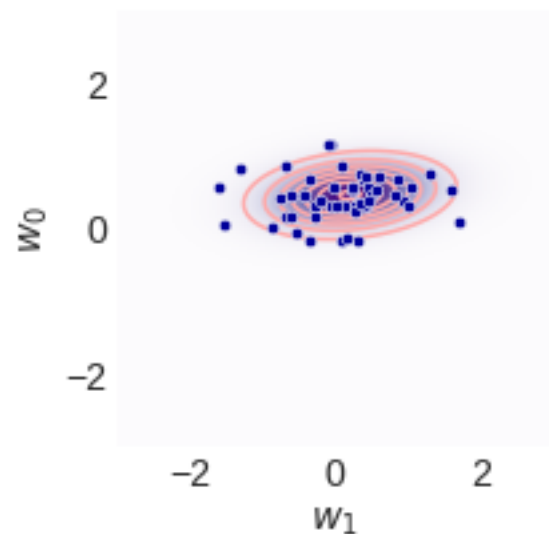
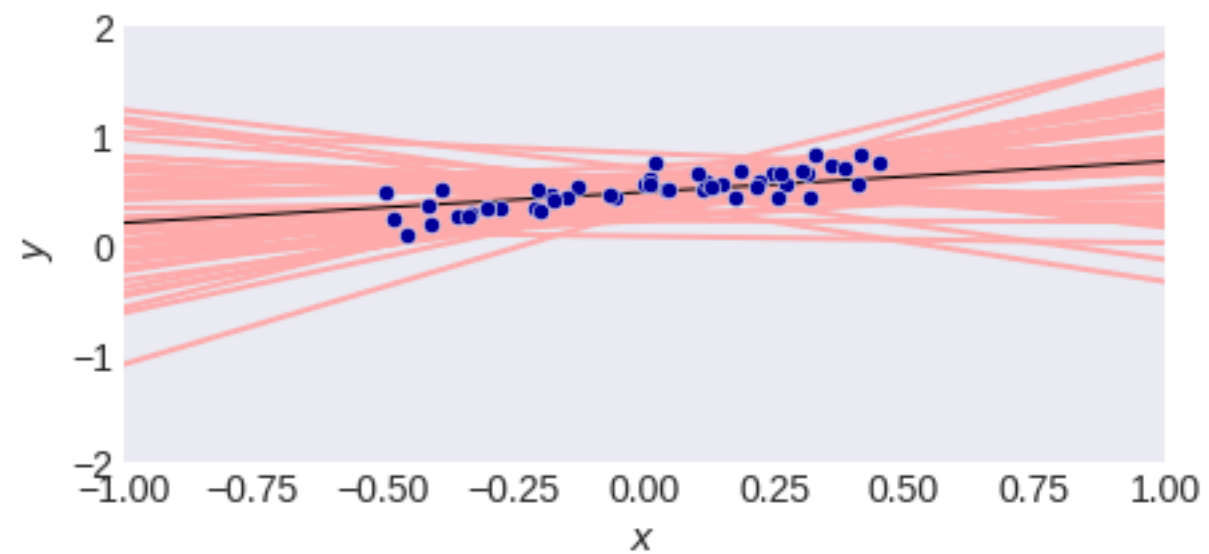
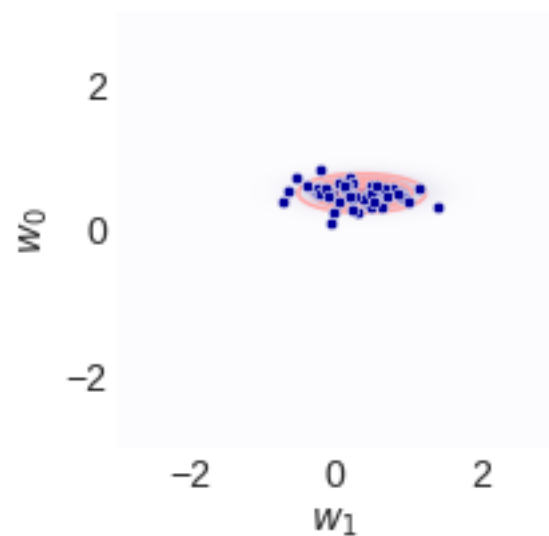


$$m = 3$$



тут, правда, ограничения на оба коэффициента (и своб. член), что непрактично

## Байесовская теория для линейной регрессии

 $m = 10$  $m = 50$ 

## Логистическая регрессия + байесовский подход

**Здесь всё аналогично...**

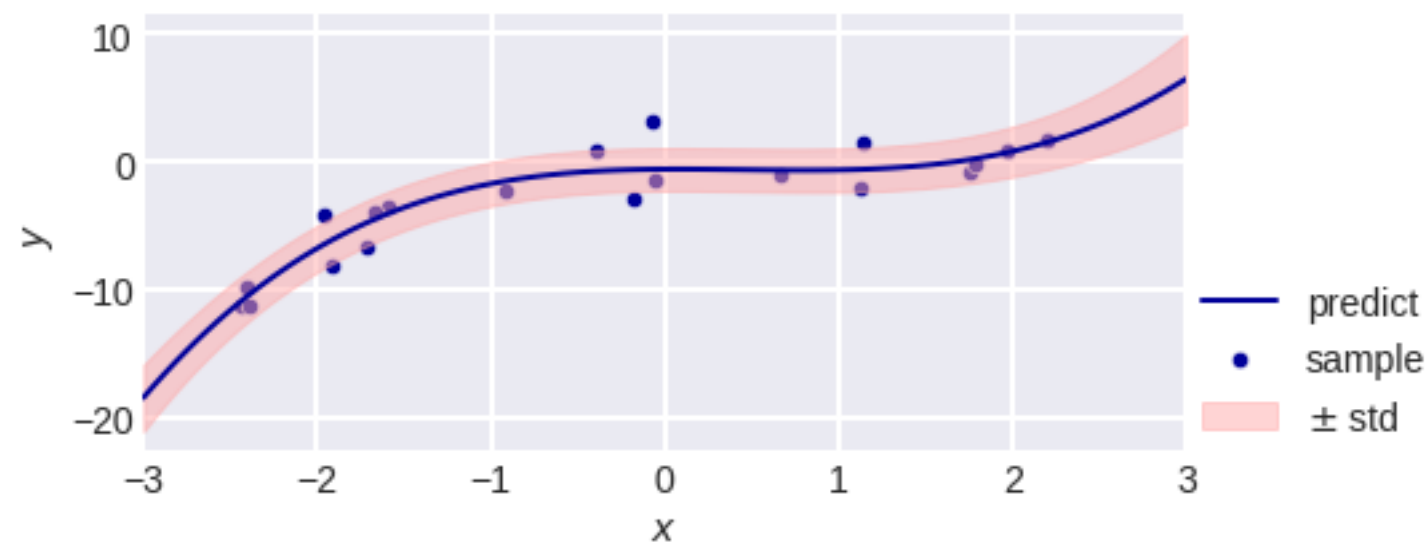
$$p(w) = \text{norm}(w \mid 0, \lambda^{-1}I) = \frac{\lambda^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{\lambda}{2} w^T w\right]$$

$$\log p(w \mid D) \sim \log p(w) + \log p(D \mid w) \sim$$

$$\sum_i (-y_i \log a_i - (1 - y_i) \log(1 - a_i)) + \lambda^2 w^T w \rightarrow \min$$

**тоже возникло регуляризационное слагаемое...**

## Минутка кода



```
from sklearn.linear_model import BayesianRidge
# обучение
model = BayesianRidge()
x = np.vander(x, 4)
model.fit(x, y)
# формируем ответ
grid = np.linspace(-3, 3, 100)
x2 = np.vander(grid, 4)
means, stds = model.predict(x2, return_std=True)
```

**Байесовские точечные оценки**  
**оценки параметра по апостериорному распределению**

**Среднее значение распределения**

$$\hat{\theta} = \int \theta p(\theta | y) d\theta$$

**очень трудоёмко!**

**Наилучшая оценка с смысле минимизации апостериорного байесовского риска:**

$$\int (\theta - \hat{\theta})^2 p(\theta | y) d\theta \rightarrow \min$$

**Модальное значение**

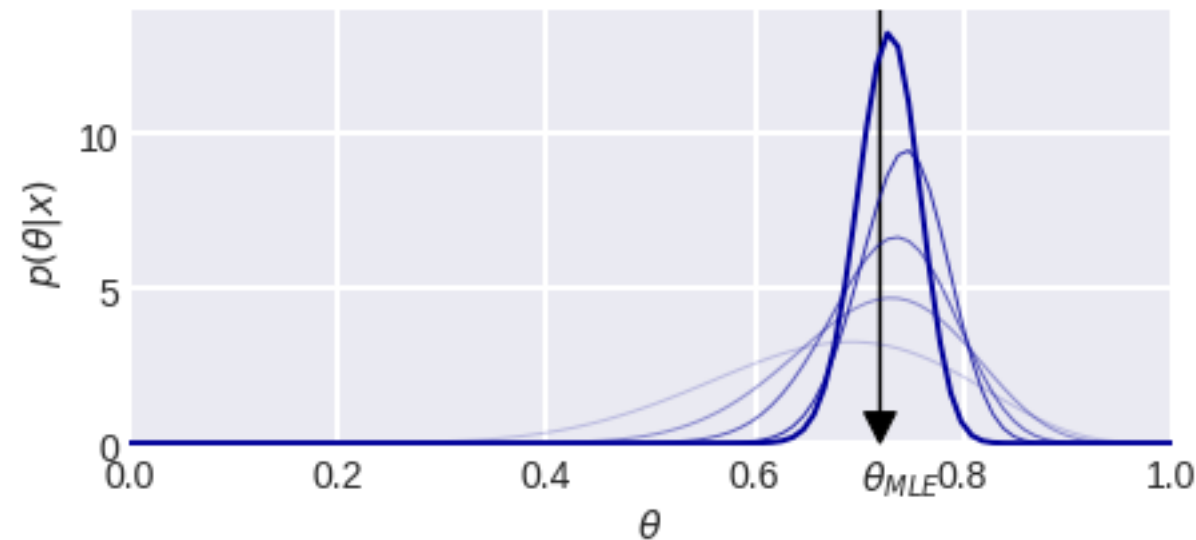
$$\hat{\theta}_{\text{MAP}} = \arg \max p(\theta | y)$$

**получается коррекция MLE (ММП):**

$$\arg \max (p(y | \theta) + p(\theta))$$

## Байесовские точечные оценки

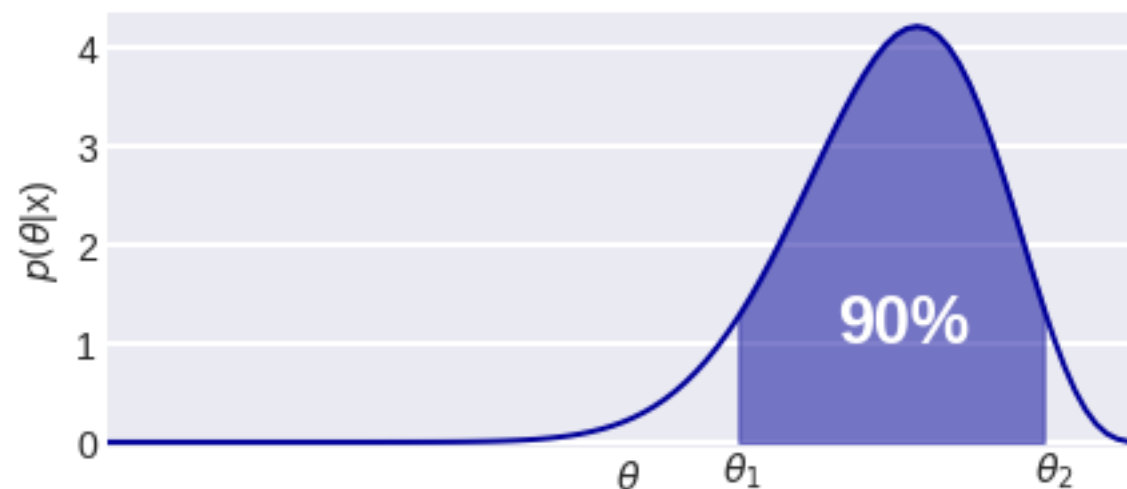
$$\lim_{m/n \rightarrow \infty} p(\theta | y_1, \dots, y_m) \rightarrow \sigma(\theta - \theta_{\text{MLE}})$$



**Медианное значение – используется реже...**

## Байесовские интервальные оценки

Из вычисленного распределения  $p(\theta | y)$



Ищем пороги  $p_1, p_2$ :

$$\int_{p_1}^{p_2} p(\theta | y) d\theta = p$$



## Проблемы с вычислениями

$$p(\theta | D) = \frac{p(D | \theta) \cdot p(\theta)}{\int_{\theta} p(D | \theta) p(\theta) d\theta}$$

**Как вычислить интеграл в знаменателе?!**

**Интегрирование в многомерном пространстве...**

**Нам повезло (сопряжённые распределения) и ответ получался даже аналитически**

**Hint:** в качестве априорного брать распределение сопряжённое к правдоподобию  
(и так формализовать начальные знания)

**Для распределений из экспоненциального семейства существуют сопряжённые**

## Как считать нормировочную константу

$$p(w | y) = \frac{p(y | w) \cdot p(w)}{p(y)} = \frac{p(y | w) \cdot p(w)}{\int p(y | w) p(w) \partial w}$$

для этого используется т.н. **Variational Bayesian Inference**:

$$\begin{aligned} \log p(y) &= \int q(z) \log p(y) \partial z = \int q(z) \log \frac{p(y, z)}{p(z | y)} \partial z = \\ &= \int q(z) \log \frac{p(y, z) q(z)}{q(z) p(z | y)} \partial z = \\ &= \underbrace{\int q(z) \log \frac{p(y, z)}{q(z)} \partial z}_{f(q) \rightarrow \max} + \underbrace{\int q(z) \log \frac{q(z)}{p(z | y)} \partial z}_{D_{\text{KL}}(p(z|y) \| q(z)) \geq 0} \end{aligned}$$

левая часть не зависит от  $q$ , поэтому, когда  $f$  максимально  $DL=0$

приближённое вычисление  $p(y)$  свели к максимизации интеграла

## Байесовский подход к машинному обучению

**обучающая выборка:**  $(X, Y)$

**обучение:**

$$p(w | X, Y) = \frac{p(Y | X, w) p(w)}{\int p(Y | X, w) p(w) \partial w}$$

**Получаем распределение в пространстве параметров модели,  
т.е. «вероятностный ансамбль»!**

**На контрольной выборке  $x'$ :**

$$p(y' | x', X, Y) = \int p(y' | x', w) p(w | X, Y) \partial w$$

**Предсказание с использованием полного распределения!**

**В этом и есть защита от переобучения...**

**ясно как дообучить модель в рамках байесовского подхода...**

**Но как делать интегрирования на этом слайде?**

## Байесовский подход к машинному обучению

**Если**

$$w_{\text{MAP}} = \arg \max p(w | X, Y) = \arg \max p(Y | X, w) p(w)$$

**теперь функцию внутри интеграла заменить на дельта-функцию:**

$$p(y' | x', X, Y) = \int p(y' | x', w) p(w | X, Y) \partial w \approx p(y' | x', w_{\text{MAP}})$$

**Мы уже видели:**

**БП  $\Rightarrow$  регуляризация  
в БП можно обучать батчами**

**Теперь понимаем:  
автоматическое ансамблирование  
есть возможность автоматического получения наиболее простой модели**

## Иерархические модели

Пусть  $w$  зависит от другого параметра (мета-параметра)  $\alpha$ ,  
тогда по формуле Байеса

$$p(\alpha | X, Y) = \frac{p(Y | X, \alpha) p(\alpha)}{p(Y | X)}$$

Правдоподобие  $\alpha$

$$p(Y | X, \alpha) = \int p(Y | X, w) p(w | \alpha) \partial w$$

называется **обоснованностью**  
(просто исключаем переменную)

– это принцип максимальной обоснованности  
параметры  $\alpha$  выбираются так,  
чтобы создать ограничения на параметр  $w$

Не пытаемся оптимизировать сразу по всем параметрам

## Иерархические модели

**Параметр  $\alpha$  может зависеть от другого параметра  
(мета-мета-параметра)...**

**считается, что при настройке таких параметров меньше риска переобучения**

### Принцип максимальной обоснованности

$$\alpha_{\text{ME}} = \arg \max p(Y | X, \alpha)$$

**далее находим апостериорное распределение**

$$p(w | \alpha_{\text{ME}}, X, Y)$$

**Так производится классификация / регрессия**

$$p(y' | x', X, Y) \approx \int p(y' | x', w) p(w | \alpha_{\text{ME}}, X, Y) \partial w$$

**RVM (метод релевантных векторов) для регрессии**

**напомним, что для линейной регрессии**  $y = w^T x + \varepsilon$ ,  $\varepsilon \sim \text{norm}(0, \sigma^2)$

$$p(w) = \text{norm}(w | 0, \lambda^{-1} I) \rightarrow \frac{1}{2\sigma^2} \sum_{i=1}^m (y_i - w^T x_i)^2 + \frac{\lambda}{2} w^T w \rightarrow \min$$

**пусть теперь**

**априорное распределение на параметры**

$$p(w | \alpha) = \text{norm}(w | 0, \text{diag}(\alpha_1, \dots, \alpha_n)^{-1})$$

**для каждого параметра независимая регуляризация**

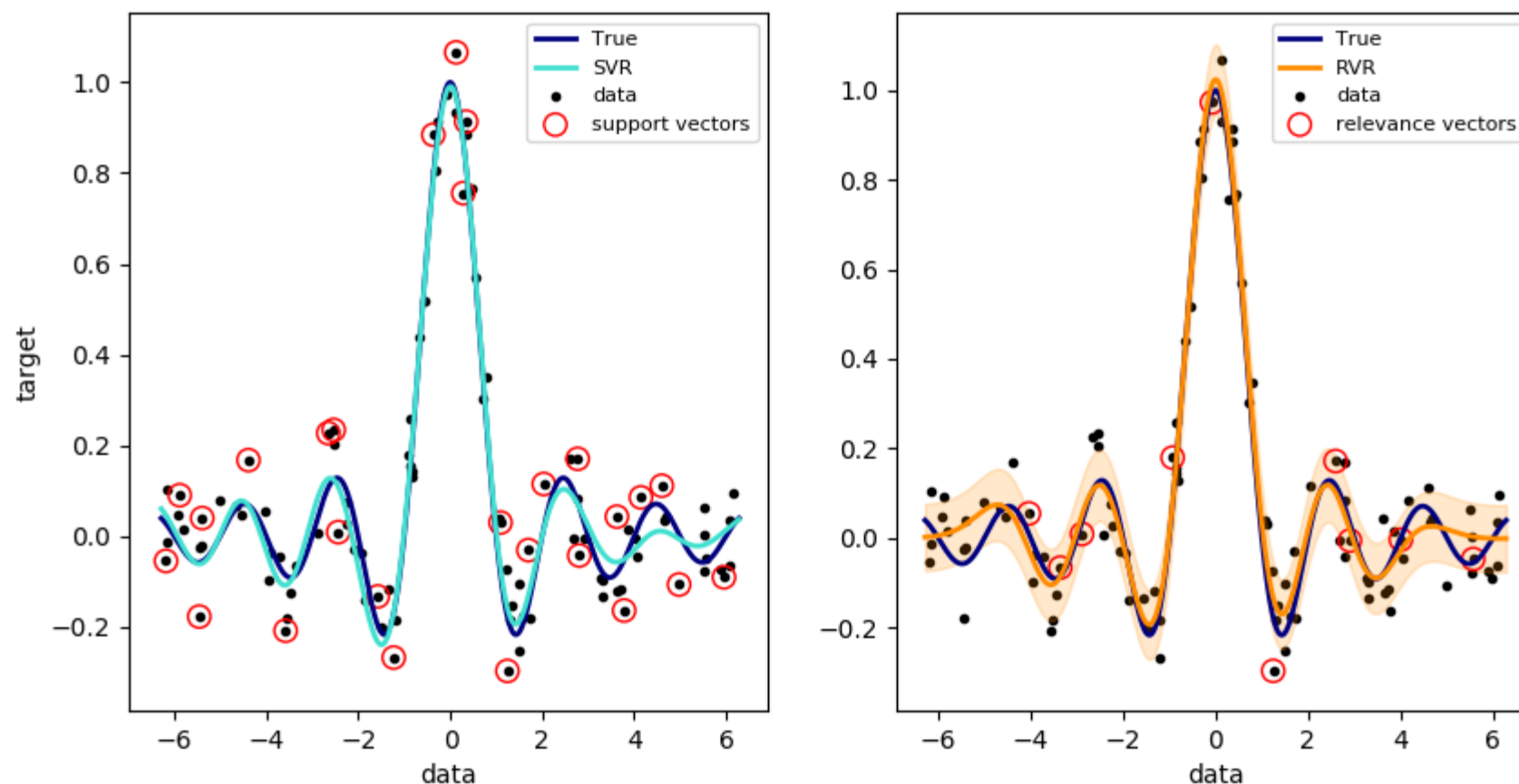
**по максимальной обоснованности**

$$p(Y | X, \alpha, \sigma^2) = \int p(Y | X, w, \sigma^2) p(w | \alpha) \partial w \rightarrow \max_{\alpha, \sigma^2}$$

**вывод не рассматриваем**

# RVM (метод релевантных векторов) для регрессии

RVR versus SVR



**параметры  $\alpha, \sigma^2$  подбираются автоматически**  
**разреженное решение – «отбор релевантных векторов»**  
**кроме прогноза выдаёт дисперсию прогноза**

[https://sklearn-rvm.readthedocs.io/en/latest/auto\\_examples/plot\\_compare\\_rvr\\_svr.html](https://sklearn-rvm.readthedocs.io/en/latest/auto_examples/plot_compare_rvr_svr.html)



**RVM (метод релевантных векторов) для классификации**

**напомним, что для логистической регрессии**

$$p(w) = \text{norm}(w \mid 0, \lambda^{-1}I) \rightarrow \text{logloss} + \lambda w^T w \rightarrow \min$$

**пусть теперь**

**априорное распределение на параметры**

$$p(w \mid \alpha) = \text{norm}(w \mid 0, \text{diag}(\alpha_1, \dots, \alpha_n)^{-1})$$

**для каждого параметра независимая регуляризация**

**и опять принцип максимальной обоснованности**

**тут ещё приём – приближение Лапласа, чтобы взять интеграл**

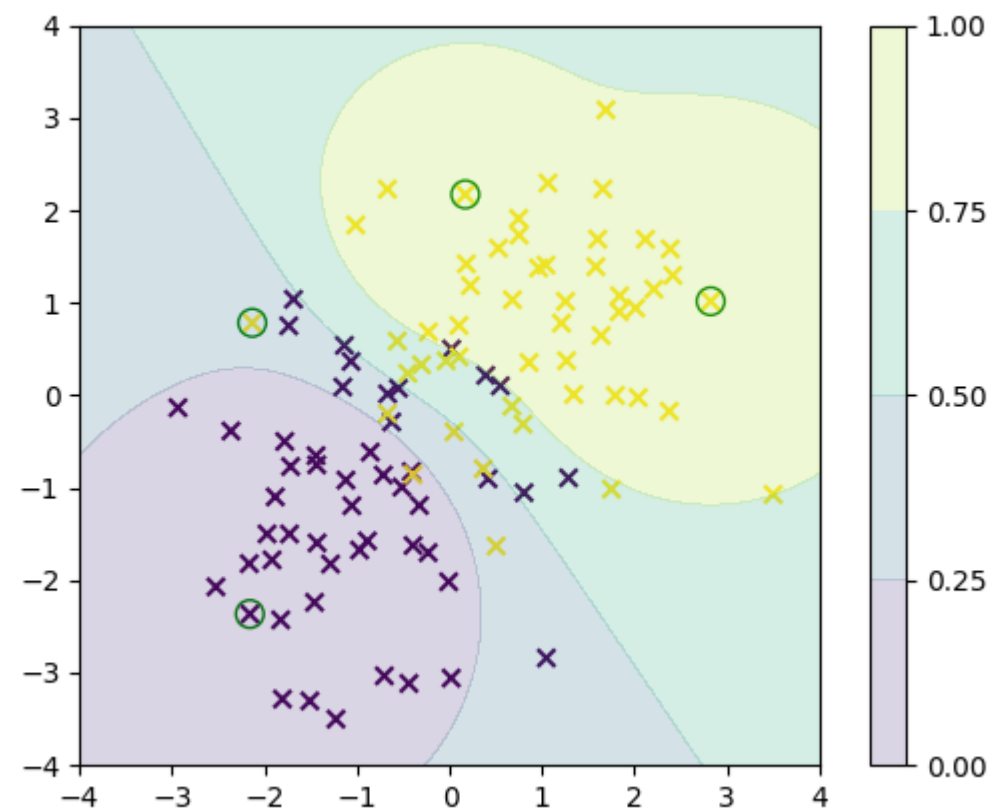
**SVM**

$$\sum_i \max[1 - y_i w^T x, 0] + \alpha w^T w \rightarrow \min$$

**RVM**

$$\sum_i \log(1 + \exp(-y_i w^T x)) + w^T \text{diag}(\alpha) w \rightarrow \min$$

## RVM (метод релевантных векторов) для классификации



**опять автоматический подбор параметров**  
**разреженное решение**  
**не рассматриваем вывод**

[https://sklearn-rvm.readthedocs.io/en/latest/auto\\_examples/plot\\_rvm\\_for\\_classification.html](https://sklearn-rvm.readthedocs.io/en/latest/auto_examples/plot_rvm_for_classification.html)

## Что такое случайность

Частотный подход (frequentism)	Байесовский подход (Bayesianism)
– объективная неопределённость	– мера нашего незнания
=> делаем серию экспериментов (в пределе истина)	переход от априорной информации (формализует незнание) к апостериорной (учитывая эксперимент)
параметры делятся на детерминированные и случайные	ВСЕ параметры – случайные (они же неизвестны!)
Частота связана с вероятностью! Если событие не повторяется, то говорить о вероятности бессмысленно	Можно говорить о любых событиях, например, о вероятности поражения в финале ЧМ.
Типичный метод – MLE (доказана оптимальность) обычно несмещённая оценка, но большой разброс	Формула Байеса + другие ф-лы MAP – обычно смещённая, но малый разброс
хорошо при $n \rightarrow \infty$	Что-то разумное при малых $n$ при $n=0$ априорное распределение = апостериорному

## Bayesian learning

- обучает целое распределение параметров
- более робастное решение за счёт постериорного усреднения (posterior averaging)
- позволяет оценить **достоверность/уверенность** (confidence) предсказания модели
  - позволяет обучать гиперпараметры по данным
  - позволяет обучаться с пропущенными значениями
  - связь с логическим выводом (Modus Ponens)

$$\frac{A, A \rightarrow B}{A \& B}$$

$$\frac{p(A), p(B | A)}{p(A \& B)}$$

## Камни в огород Байесианцев

**Можно поверить в распределения объектов**  
(в конце концов, мы хотим узнать, как пространство объектов устроено),  
**но параметры модели...**

**Большинство примеров «искусственные» – для хороших распределений**

**Утверждается, что байесовский подход работает даже при отсутствии выборки (экспериментов). Тогда априорное распределение = апостериорному...  
но откуда взять априорное?!**

Удобно вносить свои знания с помощью распределения,  
но можем внести и заблуждения...

**Методы вычислительно ресурсоёмкие**

## Оптимизационный подход к МО

- **выбрать модель** (предполагаем, что достаточно хорошо описывает данные)
- **выбрать метрику качества** (которая формализует «хорошесть» решения)
  - **оптимизировать метрику** варьируя параметры модели

## Вероятностный подход к МО

- **выбрать вероятностную модель данных**  
(предполагаем, что достаточно хорошо описывает данные)
- **в вероятностных терминах сформулировать «хорошесть» решения** (MLE или MAP)
  - **оптимизировать вероятностный критерий** (MLE или MAP)

## Источники

**Bishop C. M. Pattern recognition and machine learning. – Springer, 2006**

**Буре В.М., Грауэр Л.В. Лекция «Байесовский подход» // ШАД СПб, 2013**

**[https://compscicenter.ru/media/slides/math\\_stat\\_2013\\_spring/2013\\_05\\_22\\_math\\_stat\\_2013\\_spring.pdf](https://compscicenter.ru/media/slides/math_stat_2013_spring/2013_05_22_math_stat_2013_spring.pdf)**

**Материалы DeepBayes-2017 // <http://deepbayes.ru/2017/>**

**Учебное пособие по Байесовскому подходу**

**<http://www.machinelearning.ru/wiki/images/4/43/BayesML-2007-textbook-2.pdf>**

**Байесовский подход**

**<https://dyakonov.org/2018/07/30/байесовский-подход/>**

**Сайт Типпинга**

**<http://www.miketipping.com/sparsebayes.htm>**