Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н., н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

Задачи классификации
(продолжение)

Михаил Криницкий

к.т.н., н.с.
Институт океанологии РАН им. П.П. Ширшова

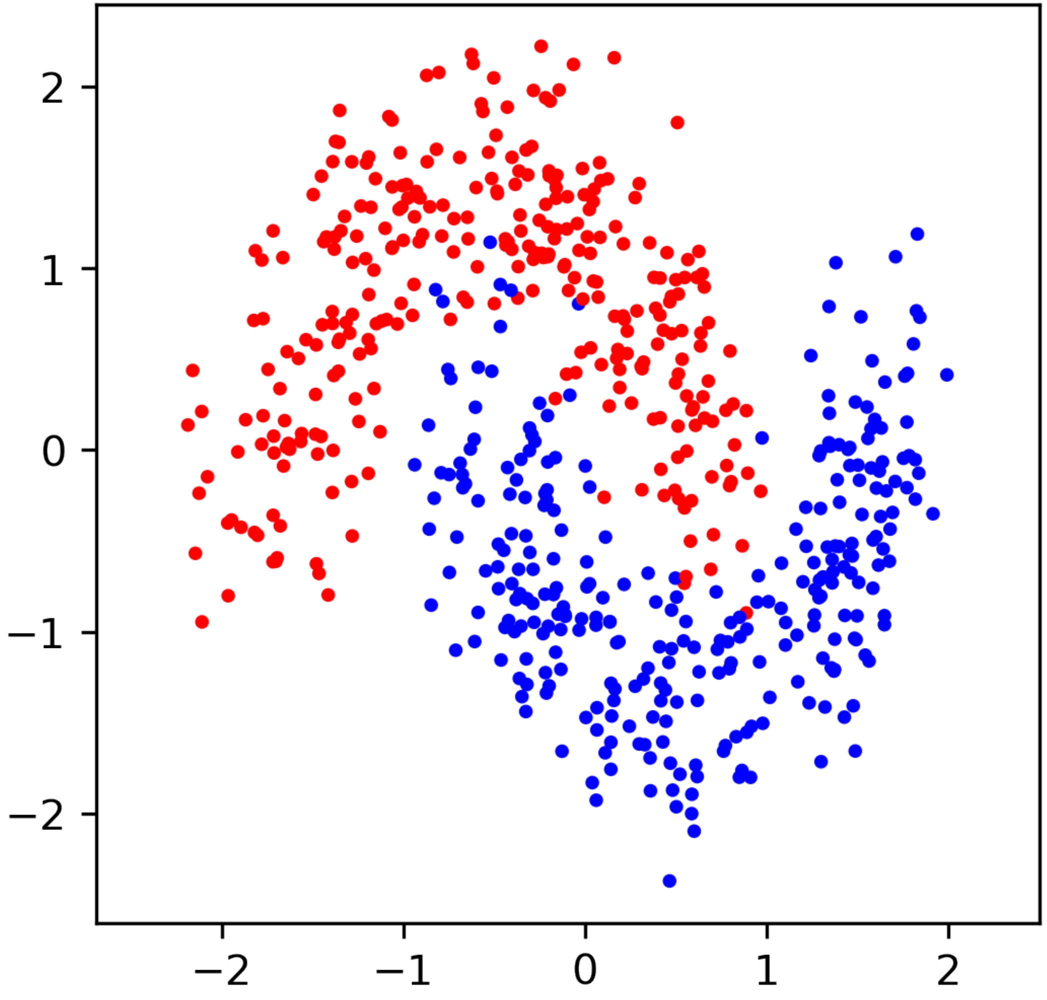
Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

previously on ML4ES

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»
  - восстановление регрессии
  - классификация



что я хочу? – метку класса

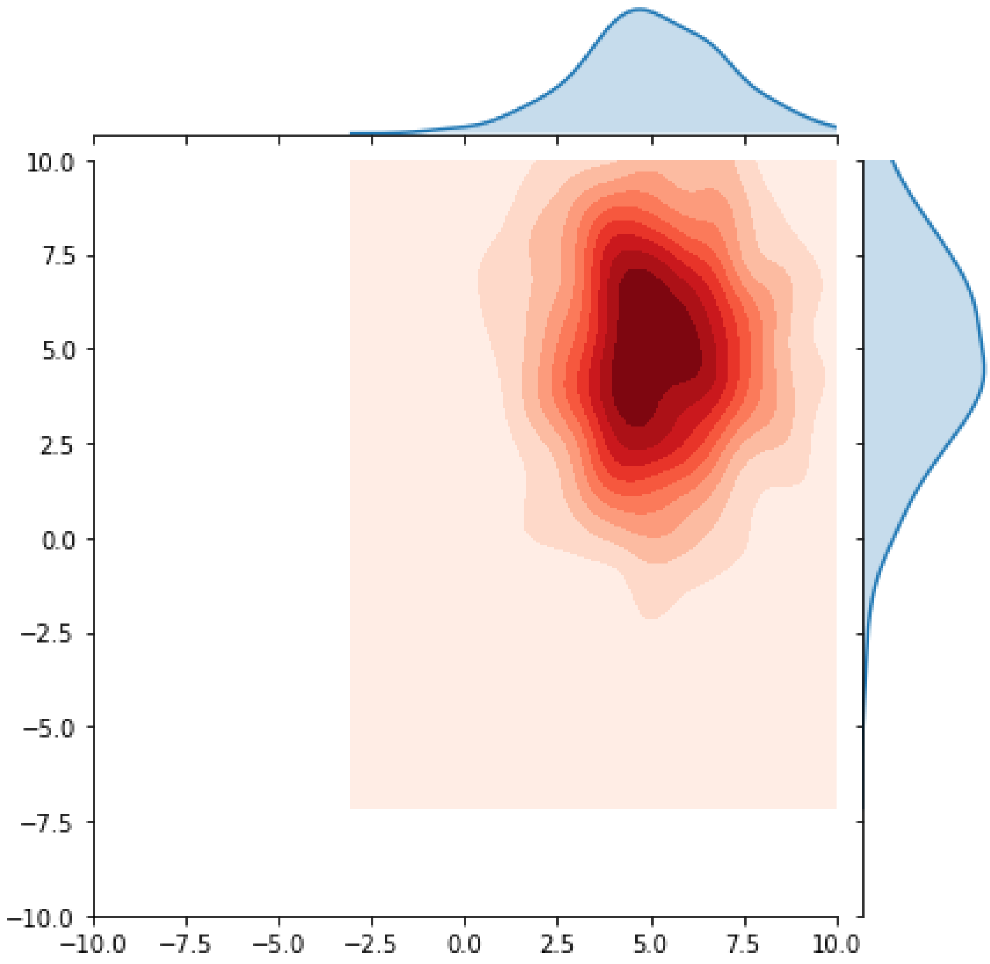
«красный или синий?»

(бинарная классификация)

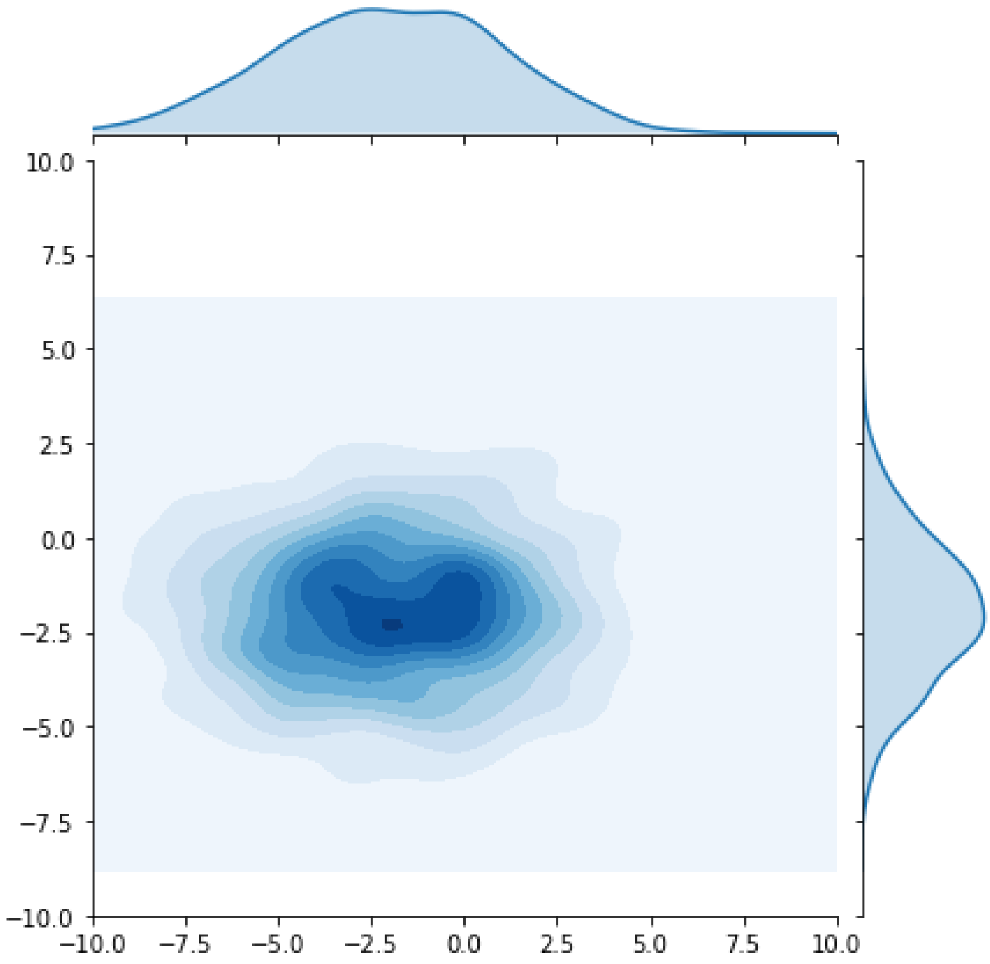
ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Цель: оценить **вероятность** классов $\boldsymbol{A}$ (класс "**1**") и $\boldsymbol{B}$ (класс "**0**") для объекта, описываемого значением x.

$$P(Y = k | X = x)$$



$Y = A$

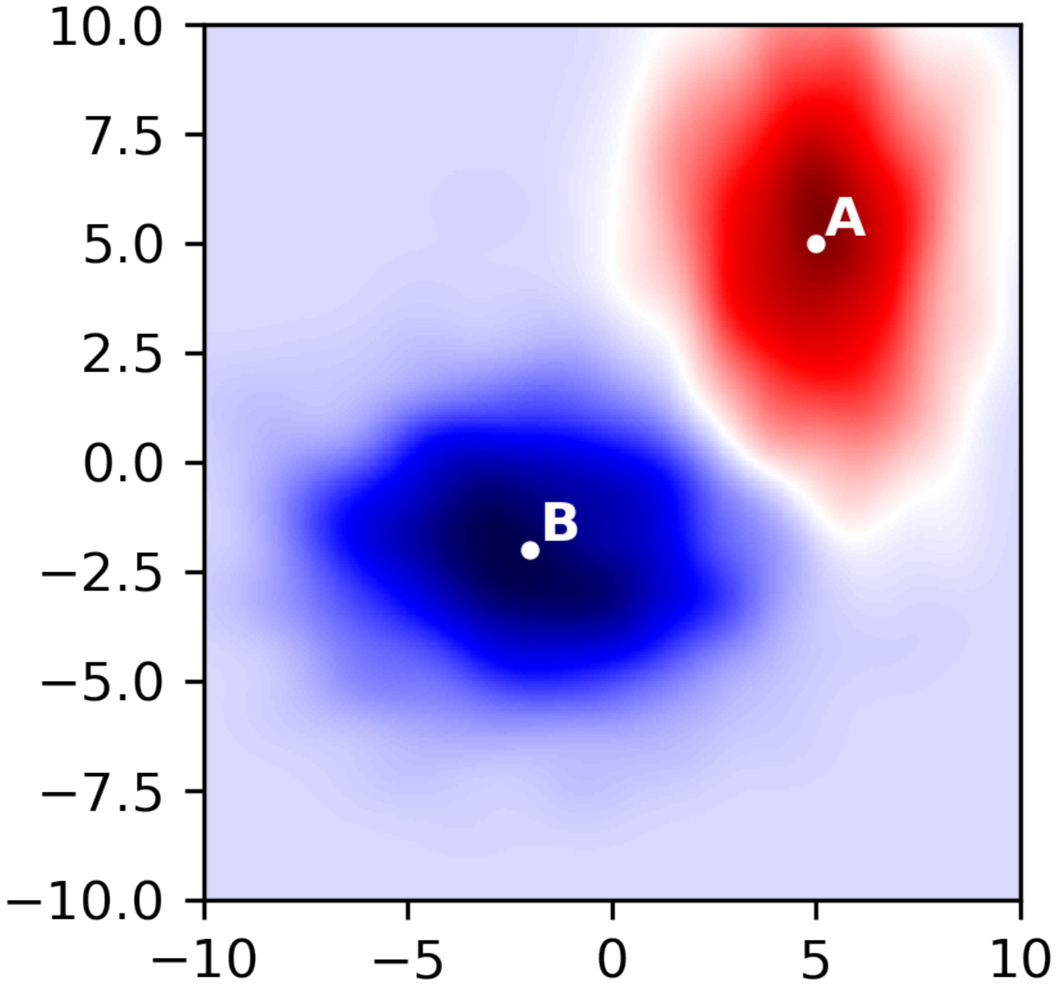


$Y = B$

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Цель: оценить **вероятность** классов A (класс “**1**”) и B(класс “**0**”) для объекта, описываемого значением x.

$$P(Y = k|X = x)$$



ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Naïve bayes classifier, «наивный байесовский классификатор»

$$P(Y|X) \propto P(X|Y)\ P(Y)$$

можно забыть про $P(X = x)$

$$P(X|Y) = P\left(x_1, x_2, x_3 \dots x_p \middle| Y\right) = P\left(x_2, x_3 \dots x_p \middle| Y, x_1\right) * P(x_1|Y) = \cdots$$

$$= P(x_1|Y) * P(x_2|Y, x_1) * P(x_3|Y, x_1, x_2) * \cdots * P(x_p|Y, x_1, \dots, x_{p-1})$$

наивное предположение: переменные x_i условно независимы:

$$P\left(x_p \middle| Y, x_1, x_2, \dots, x_{p-1}\right) = P\left(x_p \middle| Y\right)$$

Остается оценить распределения $P(x_k|Y)$ для всех k независимо друг от друга – и можно подставлять в оценку вероятности $P(Y|X)$. Оценка распределения может базироваться на предположении об их нормальности, - тогда нужно оценить выборочное среднее (оценка параметра μ) и выборочную дисперсию (оценка параметра σ^2).

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Naïve bayes classifier, «наивный байесовский классификатор»

$$P(Y|X) \propto P(X|Y)\ P(Y)$$

можно забыть про $P(X = x)$

$$P(X|Y) = P(x_1, x_2, x_3 \dots x_p|Y) = P(x_2, x_3 \dots x_p|Y, x_1) * P(x_1|Y) = \cdots$$

$$= P(x_1|Y) * P(x_2|Y, x_1) * P(x_3|Y, x_1, x_2) * \cdots * P(x_p|Y, x_1, \dots, x_{p-1})$$

наивное предположение: переменные x_i условно независимы:

$$P(x_p|Y, x_1, x_2, \dots, x_{p-1}) = P(x_p|Y)$$

Остается оценить распределения $P(x_k|Y)$ для всех k независимо друг от друга – и можно подставлять в оценку вероятности $P(Y|X)$. Оценка распределения может базироваться на предположении об их нормальности, - тогда нужно оценить выборочное среднее (оценка параметра μ) и выборочную дисперсию (оценка параметра σ^2).

NOTE!

LDA (Linear Discriminant Analysis, метод линейного дискриминантного анализа) – то же самое, но в предположении, что (а) распределение $P(x_p|Y)$ - **нормальное**; (б) для всех компонент x_i **дисперсия одинакова**.

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Naïve bayes classifier, «наивный байесовский классификатор»

$$P(Y|X) \propto P(X|Y)\ P(Y)$$

можно забыть про $P(X = x)$

$$P(X|Y) = P\left(x_1, x_2, x_3 \ldots x_p \middle| Y\right) = P\left(x_2, x_3 \ldots x_p \middle| Y, x_1\right) * P(x_1|Y) = \cdots$$

$$= P(x_1|Y) * P(x_2|Y, x_1) * P(x_3|Y, x_1, x_2) * \cdots * P(x_p|Y, x_1, \ldots, x_{p-1})$$

наивное предположение: переменные x_i условно независимы:

$$P\left(x_p \middle| Y, x_1, x_2, \ldots, x_{p-1}\right) = P\left(x_p \middle| Y\right)$$

Остается оценить распределения $P(x_k|Y)$ для всех k независимо друг от друга – и можно подставлять в оценку вероятности $P(Y|X)$. Оценка распределения может базироваться на предположении об их нормальности, - тогда нужно оценить выборочное среднее (оценка параметра μ) и выборочную дисперсию (оценка параметра σ^2).

NOTE!

QDA (Quadratic Discriminant Analysis, метод квадратичного дискриминантного анализа) – то же самое, но в предположениях, что (а) распределение $P\left(x_p \middle| Y\right)$ - **нормальное**; (б) для всех компонент x_i **дисперсии разные** (и здесь σ^2 становится матрицей ковариаций Σ^2). NB с нормальным распределением $P(x_p|Y)$ - это QDA с диагональной Σ^2

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Naïve bayes classifier, «наивный байесовский классификатор»

$$P(Y|X) \propto P(X|Y)\ P(Y)$$

Наивное предположение об условной независимости предикторов

$$P\left(x_p \middle| Y = k, x_1, x_2, \dots, x_{p-1}\right) = P\left(x_p \middle| Y = k\right)$$

Оценка параметров распределений (отдельно для каждого класса, для каждой компоненты признакового описания), подстановка в ф-лу Байеса

$$P(Y = k|X = x) \propto P(Y) * \prod_{j=1}^{p} P\left(x_j \middle| Y = k\right)$$

Вычисление вывода модели

$$\hat{y}(x) = \mathrm{argmax}(P_k(x))$$

Нормировка вероятностей (если нужны оценки вероятностей)

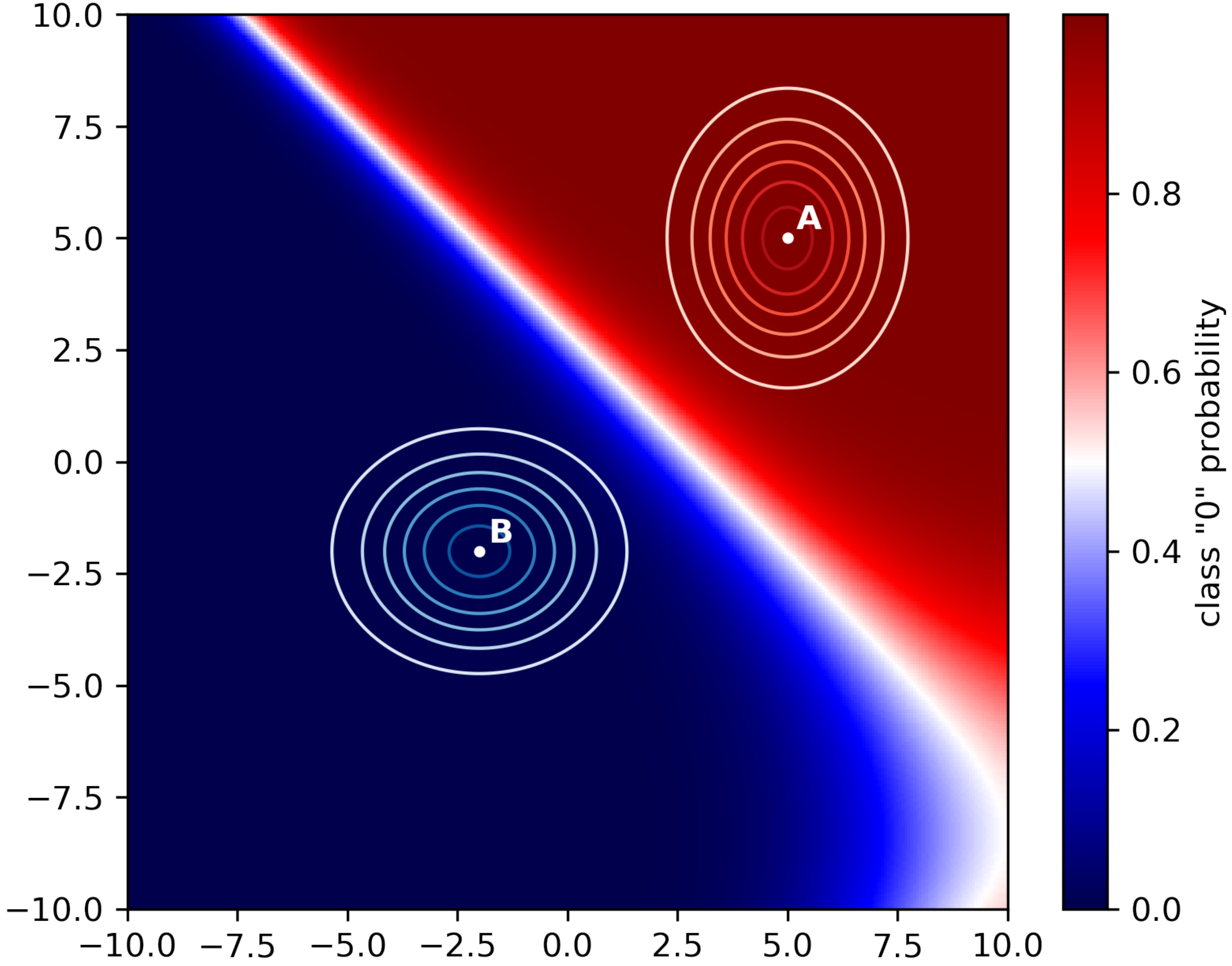
$$P_k(x) = \frac{P(Y = k|X = x)}{\sum_k P(Y = k|X = x)}$$

code demonstration

previously on ML4ES

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

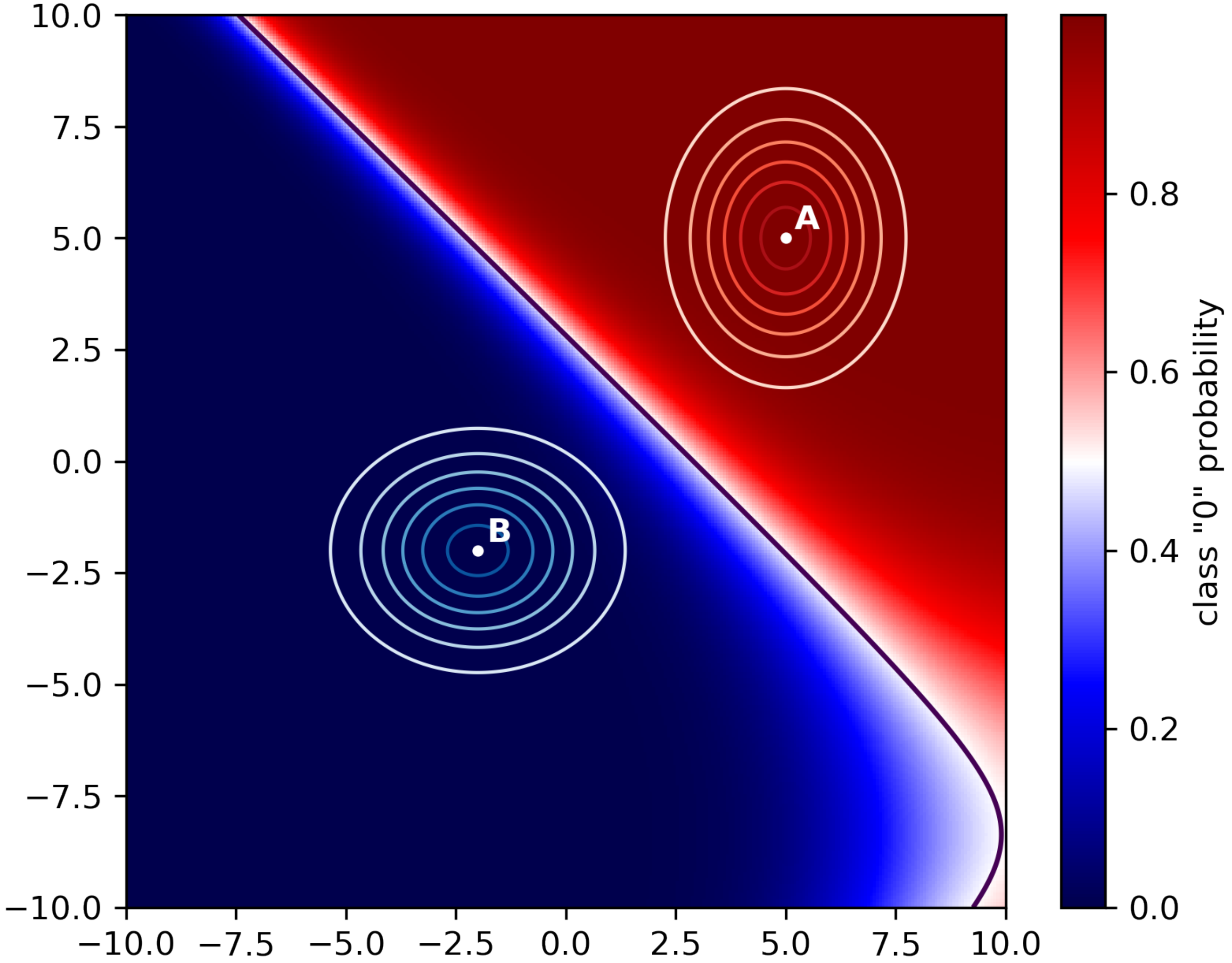
Naïve bayes classifier, «наивный байесовский классификатор»



previously on ML4ES

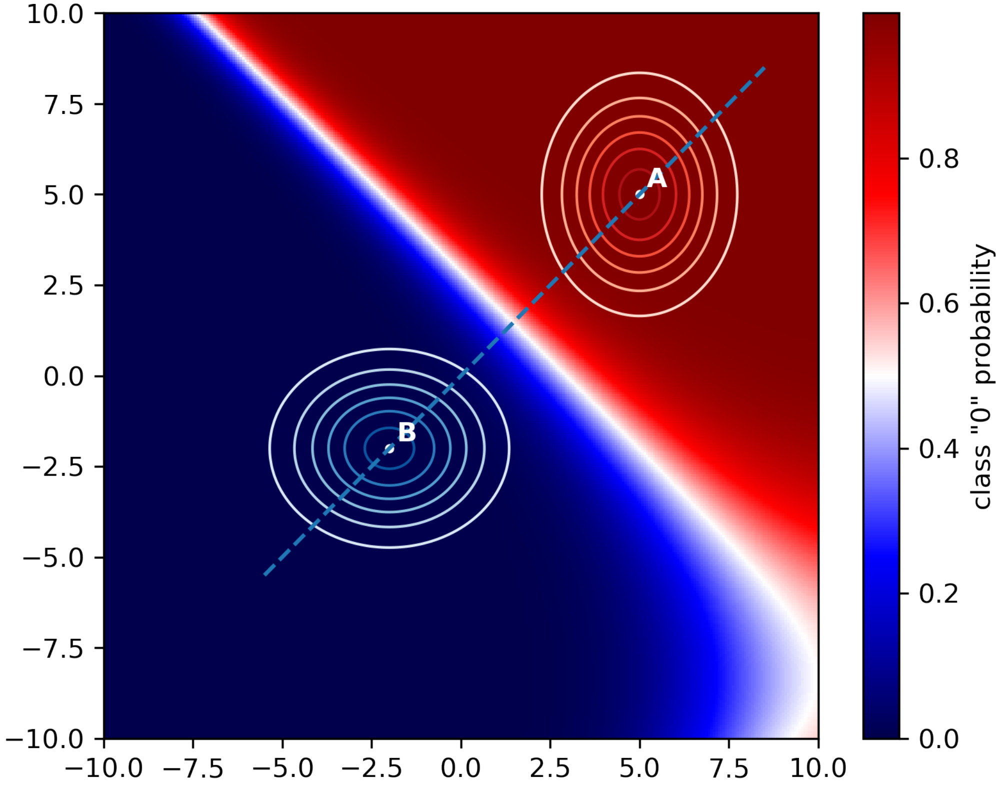
ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

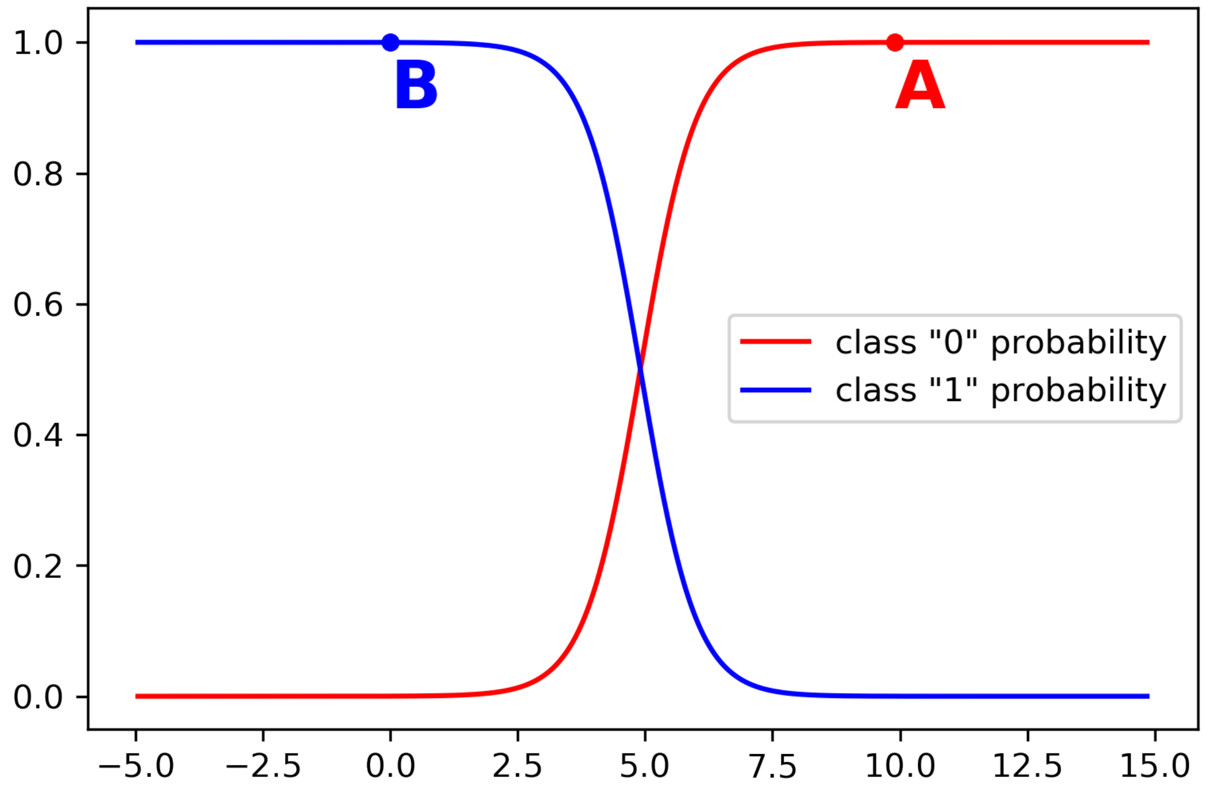
Разделяющая поверхность
(на примере результатов наивного байесовского классификатора)



previously on ML4ES

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации





Может быть, можно аппроксимировать $P(y = 0|x)$ на основании данных обучающей выборки?
Разве не для этого придуман весь подход машинного обучения?

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Логистическая регрессия («logistic regression»)

previously on ML4ES

Линейная регрессия: набор предположений “LINE”

$$y_i \sim \mathcal{N}(\theta \cdot x_i, \sigma^2)$$

$$L(\mathcal{T}) = P(\{y_i, x_i\}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta \cdot x_i)^2}{2\sigma^2}}$$

$$\ell(\mathcal{T}, \theta) = \log L(\mathcal{T}, \theta) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{N} \frac{(y_i - \theta \cdot x_i)^2}{2\sigma^2}$$

$$\theta^* = \underset{\Theta}{\operatorname{argmax}}\big(\ell(\mathcal{T}, \theta)\big) = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{(y_i - \theta \cdot x_i)^2}{2\sigma^2}$$

Логистическая регрессия:

(1) Предполагаем, что переменная Y распределена согласно распр.-ю Бернулли с параметром p (зависящий от x_i);
(2) Предполагаем, что отношение вероятностей классов «1» и «0» соотносятся как линейная функция $f(\theta, x) = \theta \cdot x$

$$Y \sim \mathcal{B}\big(p(\theta, x)\big), \qquad \log \frac{p(\theta,x)}{1 - p(\theta,x)} = \theta \cdot x$$

$$p(\theta, x) = \frac{1}{1 + e^{-\theta \cdot x}}$$

$$L(\mathcal{T}) = \prod_{i=1}^{N} \left(p(x_i)^{y_i} * \big(1 - p(x_i)\big)^{(1 - y_i)}\right)$$

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Логистическая регрессия («logistic regression»)

$\log\frac{p(\theta,x)}{1-p(\theta,x)} = \theta \cdot x, \qquad p(\theta,x) = \frac{1}{1+e^{-\theta \cdot x}}$

Новое обозначение (для краткости): $p_i \equiv p(\theta, x_i)$

$$L(\mathcal{T}) = \prod_{i=1}^{N}\left(p_i^{y_i} * (1-p_i)^{(1-y_i)}\right)$$

Бинарная кросс-энтропия (отрицательная)

Negative binary cross-entropy

$$\ell(\mathcal{T},\theta) = \log(L(\mathcal{T},\theta)) = \sum_i \log\left(p_i^{y_i}\right) + \sum_i \log\left((1-p_i)^{(1-y_i)}\right) =$$

$$= \sum_i (y_i * \log p_i + (1-y_i) * \log(1-p_i)) = \sum_i \log(1-p_i) + \sum_i y_i * \log\frac{p_i}{1-p_i} =^{(*)}$$

$$\left\{p_i = \frac{1}{1+e^{-\theta \cdot x_i}} \;:\; (1-p_i) = \frac{1}{1+e^{\theta \cdot x_i}}\right\}$$

$$\left\{\log\frac{p_i}{1-p_i} = \theta \cdot x_i\right\}$$

$$(*) = -\sum_i \log\left(1+e^{\theta \cdot x_i}\right) + \sum_i y_i * \theta \cdot x_i$$

$$\frac{\partial \ell(\mathcal{T},\theta)}{\partial \theta} = \sum_i \left(y_i - p(\theta, x_i)\right)x_i$$

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Логистическая регрессия («logistic regression»)

previously on ML4ES

Линейная регрессия:

набор предположений “LINE”

$$y_i \sim \mathcal{N}(\theta \cdot x_i, \sigma^2)$$

$$L(\mathcal{T}) = P(\{y_i, x_i\}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta \cdot x_i)^2}{2\sigma^2}}$$

$$\ell(\mathcal{T}, \theta) = \log L(\mathcal{T}, \theta) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{N} \frac{(y_i - \theta \cdot x_i)^2}{2\sigma^2}$$

$$\theta^* = \underset{\Theta}{\operatorname{argmax}}\big(\ell(\mathcal{T}, \theta)\big) = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{(y_i - \theta \cdot x_i)^2}{2\sigma^2}$$

Логистическая регрессия:

(1) Предполагаем, что переменная Y распределена согласно распределению Бернулли с параметром p;
(2) Предполагаем, что отношение вероятностей классов «1» и «0» соотносятся как линейная функция $f(\theta, x) = \theta \cdot x$

$$Y \sim \mathcal{B}\big(p(\theta, x)\big), \qquad \log \frac{p(\theta,x)}{1 - p(\theta,x)} = \theta \cdot x, \quad p(\theta, x) = \frac{1}{1 + e^{-\theta \cdot x}}$$

$$L(\mathcal{T}) = \prod_{i=1}^{N} \left(p(x_i)^{y_i} * \big(1 - p(x_i)\big)^{(1 - y_i)}\right)$$

$$\ell(\mathcal{T}, \theta) = -\sum_i \log\big(1 + e^{\theta \cdot x_i}\big) + \sum_i y_i * \theta \cdot x_i$$

$$\theta^* = \underset{\Theta}{\operatorname{argmax}}\big(\ell(\mathcal{T}, \theta)\big) = \underset{\Theta}{\operatorname{argmin}} \left(\sum_i \big(\log\big(1 + e^{\theta \cdot x_i}\big) - y_i * \theta \cdot x_i\big)\right)$$

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Логистическая регрессия («logistic regression»)

code demonstration

ОБУЧЕНИЕ С УЧИТЕЛЕМ: задача классификации

Разделяющая поверхность
(на примере результатов логистической регрессии)