Теоретические основы информатики (концептуальные модели и математические основы)

Лекция № 4. Вероятностная постановка задачи

А.С. Шундеев

Содержание

Содержание

Основные понятия (продолжение)

В дальнейшем, всегда будет предполагаться, что на множестве объектов X и на множестве меток Y заданы, некоторые σ-алгебры $\mathcal{S}_X$ и $\mathcal{S}_Y$, соответственно.

Основные понятия (продолжение)

В дальнейшем, всегда будет предполагаться, что на множестве объектов X и на множестве меток Y заданы, некоторые σ-алгебры $\mathcal{S}_X$ и $\mathcal{S}_Y$, соответственно.

Если множество меток конечно $|Y| < \infty$, то в качестве $\mathcal{S}_Y$ будет выступать 2^Y.

Основные понятия (продолжение)

В дальнейшем, всегда будет предполагаться, что на множестве объектов X и на множестве меток Y заданы, некоторые σ-алгебры $\mathcal{S}_X$ и $\mathcal{S}_Y$, соответственно.

Если множество меток конечно $|Y| < \infty$, то в качестве $\mathcal{S}_Y$ будет выступать 2^Y.

Говоря о том, что на множестве примеров Z задана вероятностная мера P, будет подразумеваться, что она задана на измеримом пространстве $(Z, \mathcal{S}_Z)$, где $\mathcal{S}_Z := \mathcal{S}_X \otimes \mathcal{S}_Y$.

Основные понятия (продолжение)

В дальнейшем, всегда будет предполагаться, что на множестве объектов X и на множестве меток Y заданы, некоторые σ-алгебры $\mathcal{S}_X$ и $\mathcal{S}_Y$, соответственно.

Если множество меток конечно $|Y| < \infty$, то в качестве $\mathcal{S}_Y$ будет выступать 2^Y.

Говоря о том, что на множестве примеров Z задана вероятностная мера P, будет подразумеваться, что она задана на измеримом пространстве $(Z, \mathcal{S}_Z)$, где $\mathcal{S}_Z := \mathcal{S}_X \otimes \mathcal{S}_Y$.

Такая вероятностная мера будет часто интерпретироваться как распределение случайного элемента (X, Y), состоящего из функций координатных проекций $X : z \longmapsto x$ и $Y : z \longmapsto y$, где $z = (x, y) \in Z$.

Основные понятия (продолжение)

Рассматриваемая функция потерь $l : Y^2 \longrightarrow [0, 1]$ будет предполагаться измеримой.

Основные понятия (продолжение)

Рассматриваемая функция потерь $l : Y^2 \longrightarrow [0, 1]$ будет предполагаться измеримой.

Будет также предполагаться, что рассматриваемый класс гипотез $\mathcal{H} \subseteq Y^X$ или класс концептов $\mathcal{C} \subseteq 2^X$ будет состоять только из измеримых отображений или, соответственно, измеримых множеств.

Основные понятия (продолжение)

Рассматриваемая функция потерь $l : Y^2 \longrightarrow [0, 1]$ будет предполагаться измеримой.

Будет также предполагаться, что рассматриваемый класс гипотез $\mathcal{H} \subseteq Y^X$ или класс концептов $\mathcal{C} \subseteq 2^X$ будет состоять только из измеримых отображений или, соответственно, измеримых множеств.

В этом случае класс функций $\mathcal{F} \subseteq [0, 1]^Z$, $\mathcal{H} \simeq_l \mathcal{F}$ будет состоять только из интегрируемых случайных величин.

Основные понятия (продолжение)

Ожидаемым риском (риском) называется функция

$$R(\mathsf{P}, l; h) := \mathop{\mathbf{E}}_{(x,y)\sim\mathsf{P}} \left[l(h(x), y) \right] \qquad (h \in \mathcal{H}),$$

Основные понятия (продолжение)

Ожидаемым риском (риском) называется функция

$$R(\mathsf{P}, l; h) := \mathop{\mathbf{E}}_{(x,y)\sim\mathsf{P}} \left[l(h(x), y) \right] \qquad (h \in \mathcal{H}),$$

а минимальным риском класса гипотез $\mathcal{H}$ называется число

$$R(\mathsf{P}, l; \mathcal{H}) := \inf_{h\in\mathcal{H}} R(\mathsf{P}, l; h).$$

Основные понятия (продолжение)

Ожидаемым риском (риском) называется функция

$$R(\mathsf{P}, l; h) := \mathop{\mathbf{E}}_{(x,y)\sim\mathsf{P}} \left[l(h(x), y) \right] \qquad (h \in \mathcal{H}),$$

а минимальным риском класса гипотез $\mathcal{H}$ называется число

$$R(\mathsf{P}, l; \mathcal{H}) := \inf_{h\in\mathcal{H}} R(\mathsf{P}, l; h).$$

Байесовской ошибкой называется величина

$$R^*(\mathsf{P}, l) := \inf_{g\in \mathcal{S}_\mathsf{X} \,|\, \mathcal{S}_\mathsf{Y}} R(\mathsf{P}, l; g),$$

Основные понятия (продолжение)

Ожидаемым риском (риском) называется функция

$$R(\mathsf{P}, l; h) := \mathop{\mathbf{E}}_{(x,y)\sim\mathsf{P}} \left[l(h(x), y) \right] \qquad (h \in \mathcal{H}),$$

а минимальным риском класса гипотез $\mathcal{H}$ называется число

$$R(\mathsf{P}, l; \mathcal{H}) := \inf_{h\in\mathcal{H}} R(\mathsf{P}, l; h).$$

Байесовской ошибкой называется величина

$$R^*(\mathsf{P}, l) := \inf_{g\in \mathcal{S}_\mathsf{X} \,|\, \mathcal{S}_\mathsf{Y}} R(\mathsf{P}, l; g),$$

при этом измеримое отображение g^*, на котором достигается этот инфимум, называется байесовским предиктором.

Основные понятия (продолжение)

Ожидаемым риском (риском) называется функция

$$R(\mathsf{P}, l; h) := \mathop{\mathbf{E}}_{(x,y)\sim\mathsf{P}} \bigl[l(h(x), y)\bigr] \qquad (h \in \mathcal{H}),$$

а минимальным риском класса гипотез $\mathcal{H}$ называется число

$$R(\mathsf{P}, l; \mathcal{H}) := \inf_{h\in\mathcal{H}} R(\mathsf{P}, l; h).$$

Байесовской ошибкой называется величина

$$R^*(\mathsf{P}, l) := \inf_{g\in\mathcal{S}_X\,|\,\mathcal{S}_Y} R(\mathsf{P}, l; g),$$

при этом измеримое отображение g^*, на котором достигается этот инфимум, называется байесовским предиктором.

Байесовскую ошибку можно интерпретировать как минимальный риск наибольшего класса гипотез $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Основные понятия (продолжение)

Избыточным риском называется величина

$$\mathcal{E}(\mathsf{P}, l; h) := R(\mathsf{P}, l; h) - R^*(\mathsf{P}, l).$$

Основные понятия (продолжение)

Избыточным риском называется величина

$$\mathcal{E}(\mathsf{P}, l; h) := R(\mathsf{P}, l; h) - R^*(\mathsf{P}, l).$$

Если из контекста понятно, о какой вероятностной мере и функции потерь идёт речь, то будут использоваться сокращения $R(h)$, $R(\mathcal{H})$, R^* и $\mathcal{E}(h)$.

Основные понятия (продолжение)

Избыточным риском называется величина

$$\mathcal{E}(\mathsf{P}, l; h) := R(\mathsf{P}, l; h) - R^*(\mathsf{P}, l).$$

Если из контекста понятно, о какой вероятностной мере и функции потерь идёт речь, то будут использоваться сокращения $R(h)$, $R(\mathcal{H})$, R^* и $\mathcal{E}(h)$.

Распространим понятия ожидаемого и минимального риска на класс функций $\mathcal{F}$. По определению положим

$$R(\mathsf{P}; f) := \mathop{\mathbf{E}}_{z \sim \mathsf{P}} \big[f(z)\big] \quad (f \in \mathcal{F}), \qquad R(\mathsf{P}; \mathcal{F}) := \inf_{f \in \mathcal{F}} R(\mathsf{P}; f).$$

Основные понятия (продолжение)

Избыточным риском называется величина

$$\mathcal{E}(\mathsf{P}, l; h) := R(\mathsf{P}, l; h) - R^*(\mathsf{P}, l).$$

Если из контекста понятно, о какой вероятностной мере и функции потерь идёт речь, то будут использоваться сокращения $R(h)$, $R(\mathcal{H})$, R^* и $\mathcal{E}(h)$.

Распространим понятия ожидаемого и минимального риска на класс функций $\mathcal{F}$. По определению положим

$$R(\mathsf{P}; f) := \mathop{\mathbf{E}}_{z \sim \mathsf{P}} \big[f(z)\big] \quad (f \in \mathcal{F}), \qquad R(\mathsf{P}; \mathcal{F}) := \inf_{f \in \mathcal{F}} R(\mathsf{P}; f).$$

Будут также использоваться сокращения $R(f)$ и $R(\mathcal{F})$.

Основные понятия (продолжение)

Обсудим введённые понятия.

Основные понятия (продолжение)

Обсудим введённые понятия.

Напомним, что в реализуемом случае задачи обучения предполагается наличие целевой функции g, и с помощью алгоритма обучения требуется построить приближение $h \approx g$, $h \in \mathcal{H}$.

Основные понятия (продолжение)

Обсудим введённые понятия.

Напомним, что в реализуемом случае задачи обучения предполагается наличие целевой функции g, и с помощью алгоритма обучения требуется построить приближение $h \approx g$, $h \in \mathcal{H}$.

Заметим, что целевая функция не обязана принадлежать классу гипотез. От неё требуется только измеримость.

Основные понятия (продолжение)

В реализуемом случае вероятностная мера P имеет следующую структуру.

Основные понятия (продолжение)

В реализуемом случае вероятностная мера P имеет следующую структуру.

На измеримом пространстве $(X, \mathcal{S}_X)$ фиксируется вероятностная мера P_X, которая отражает значимость объектов.

Основные понятия (продолжение)

В реализуемом случае вероятностная мера P имеет следующую структуру.

На измеримом пространстве $(X, \mathcal{S}_X)$ фиксируется вероятностная мера P_X, которая отражает значимость объектов.

Тогда в качестве P будет выступать распределение случайного элемента $x \longmapsto (x, g(x))$, $x \in X$.

Основные понятия (продолжение)

В реализуемом случае вероятностная мера P имеет следующую структуру.

На измеримом пространстве $(X, \mathcal{S}_X)$ фиксируется вероятностная мера P_X, которая отражает значимость объектов.

Тогда в качестве P будет выступать распределение случайного элемента $x \longmapsto (x, g(x))$, $x \in X$.

Сразу заметим, что любое измеримое множество примеров, метки которых не согласуются с целевой функцией, будет иметь нулевую P-меру.

Основные понятия (продолжение)

В реализуемом случае вероятностная мера P имеет следующую структуру.

На измеримом пространстве $(X, \mathcal{S}_X)$ фиксируется вероятностная мера P_X, которая отражает значимость объектов.

Тогда в качестве P будет выступать распределение случайного элемента $x \longmapsto (x, g(x))$, $x \in X$.

Сразу заметим, что любое измеримое множество примеров, метки которых не согласуются с целевой функцией, будет иметь нулевую P-меру.

В реализуемом случае для ожидаемого риска иногда будет удобно использовать обозначение $R(P_X, l; g, h)$.

Основные понятия (продолжение)

С помощью функции потерь оценивается ошибка предиктора на индивидуальном примере.

Основные понятия (продолжение)

С помощью функции потерь оценивается ошибка предиктора на индивидуальном примере.

По своему определению ожидаемый риск представляет собой усреднённую ошибку предиктора по всем примерам.

Основные понятия (продолжение)

С помощью функции потерь оценивается ошибка предиктора на индивидуальном примере.

По своему определению ожидаемый риск представляет собой усреднённую ошибку предиктора по всем примерам.

Наибольший вклад в неё дают наиболее значимые относительно рассматриваемой вероятностной меры P примеры.

Основные понятия (продолжение)

С помощью функции потерь оценивается ошибка предиктора на индивидуальном примере.

По своему определению ожидаемый риск представляет собой усреднённую ошибку предиктора по всем примерам.

Наибольший вклад в неё дают наиболее значимые относительно рассматриваемой вероятностной меры P примеры.

Следует отметить, что нулевое зачение риска $R(h)$ не гарантирует совпадение гипотезы h и целевой функции g.

Основные понятия (продолжение)

С помощью функции потерь оценивается ошибка предиктора на индивидуальном примере.

По своему определению ожидаемый риск представляет собой усреднённую ошибку предиктора по всем примерам.

Наибольший вклад в неё дают наиболее значимые относительно рассматриваемой вероятностной меры P примеры.

Следует отметить, что нулевое зачение риска $R(h)$ не гарантирует совпадение гипотезы h и целевой функции g.

Так в примере с коллекцией камней было построено дерево решений, которое не учитывало существование объектов с меткой **серый**.

Основные понятия (продолжение)

С помощью функции потерь оценивается ошибка предиктора на индивидуальном примере.

По своему определению ожидаемый риск представляет собой усреднённую ошибку предиктора по всем примерам.

Наибольший вклад в неё дают наиболее значимые относительно рассматриваемой вероятностной меры P примеры.

Следует отметить, что нулевое зачение риска $R(h)$ не гарантирует совпадение гипотезы h и целевой функции g.

Так в примере с коллекцией камней было построено дерево решений, которое не учитывало существование объектов с меткой **серый**.

Если множество таких объектов будет иметь нулевую P_X меру, то может произойти следующее.

Основные понятия (продолжение)

Ожидаемый риск гипотезы h, вычисляемой с помощью этого дерева, будет равен нулю.

Основные понятия (продолжение)

Ожидаемый риск гипотезы h, вычисляемой с помощью этого дерева, будет равен нулю.

Однако h и g будут отличаться на целом классе объектов с меткой **серый**, которые в силу выбора меры P_X не представляются значимыми.

Основные понятия (продолжение)

Ожидаемый риск гипотезы h, вычисляемой с помощью этого дерева, будет равен нулю.

Однако h и g будут отличаться на целом классе объектов с меткой **серый**, которые в силу выбора меры P_X не представляются значимыми.

В дальнейшем будет показано существование байесовского предиктора для случаев задач бинарной классификации и обучения функций.

Основные понятия (продолжение)

Ожидаемый риск гипотезы h, вычисляемой с помощью этого дерева, будет равен нулю.

Однако h и g будут отличаться на целом классе объектов с меткой **серый**, которые в силу выбора меры P_X не представляются значимыми.

В дальнейшем будет показано существование байесовского предиктора для случаев задач бинарной классификации и обучения функций.

Кроме того, будет показано, что он может быть явно построен.

Основные понятия (продолжение)

Ожидаемый риск гипотезы h, вычисляемой с помощью этого дерева, будет равен нулю.

Однако h и g будут отличаться на целом классе объектов с меткой **серый**, которые в силу выбора меры P_X не представляются значимыми.

В дальнейшем будет показано существование байесовского предиктора для случаев задач бинарной классификации и обучения функций.

Кроме того, будет показано, что он может быть явно построен.

К сожалению, для этого необходимо знать вероятностную меру P, которая в рассматриваемой постановке задачи обучения предполагается фиксированной, но неизвестной.

Основные понятия (продолжение)

Возникает следующий закономерный вопрос.

Основные понятия (продолжение)

Возникает следующий закономерный вопрос.

Почему в качестве класса гипотез не выбрать сразу множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$?

Основные понятия (продолжение)

Возникает следующий закономерный вопрос.

Почему в качестве класса гипотез не выбрать сразу множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$?

Этому классу заведомо принадлежит байесовский предиктор, на котором достигается байсовская ошибка.

Основные понятия (продолжение)

Возникает следующий закономерный вопрос.

Почему в качестве класса гипотез не выбрать сразу множество всех измеримых отображений $\mathcal{S}_X \mid \mathcal{S}_Y$?

Этому классу заведомо принадлежит байесовский предиктор, на котором достигается байсовская ошибка.

Она же будет служить минимальным риском для этого класса гипотез.

Основные понятия (продолжение)

Возникает следующий закономерный вопрос.

Почему в качестве класса гипотез не выбрать сразу множество всех измеримых отображений $\mathcal{S}_X \mid \mathcal{S}_Y$?

Этому классу заведомо принадлежит байесовский предиктор, на котором достигается байсовская ошибка.

Она же будет служить минимальным риском для этого класса гипотез.

Лучшего результата при выбранном способе оценки точности гипотез через их ожидаемый риск получить всё равно не удастся.

Основные понятия (продолжение)

Основная причина отказа от рассмотрения класса гипотез, у которого минимальный риск хорошо приближает байесовскую ошибку, связано с существованием такого явления, как переобучение.

Основные понятия (продолжение)

Основная причина отказа от рассмотрения класса гипотез, у которого минимальный риск хорошо приближает байесовскую ошибку, связано с существованием такого явления, как переобучение.

В следующем разделе будет представлен пример метода минимизации эмпирического риска, который всегда выбирает гипотезы с нулевым эмпирическим риском и ожидаемым риском равным $\frac{1}{2}$.

Основные понятия (продолжение)

Основная причина отказа от рассмотрения класса гипотез, у которого минимальный риск хорошо приближает байесовскую ошибку, связано с существованием такого явления, как переобучение.

В следующем разделе будет представлен пример метода минимизации эмпирического риска, который всегда выбирает гипотезы с нулевым эмпирическим риском и ожидаемым риском равным $\frac{1}{2}$.

Избыточный риск представляет собой наиболее удачную числовую характеристику точности индивидуальной гипотезы, а также ошибки обучения применительно к результату работы алгоритма обучения.

Основные понятия (продолжение)

Основная причина отказа от рассмотрения класса гипотез, у которого минимальный риск хорошо приближает байесовскую ошибку, связано с существованием такого явления, как переобучение.

В следующем разделе будет представлен пример метода минимизации эмпирического риска, который всегда выбирает гипотезы с нулевым эмпирическим риском и ожидаемым риском равным $\frac{1}{2}$.

Избыточный риск представляет собой наиболее удачную числовую характеристику точности индивидуальной гипотезы, а также ошибки обучения применительно к результату работы алгоритма обучения.

Он всегда неотрицателен, а его нулевое значение соответствует достижению оптимального результата.

Содержание

Содержание

Декомпозиция ошибки обучения

Существование байесовского предиктора

Пусть $P \in \mathcal{M}^1_+(Z)$.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Пусть $P \in \mathcal{M}^1_+(Z)$.

Будем использовать следующий факт.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Пусть $\mathsf{P} \in \mathcal{M}^1_+(Z)$.

Будем использовать следующий факт.

Если $(Y, \mathcal{S}_Y)$ является стандартным измеримым пространством, то в силу теоремы о дезинтеграции имеет место представление $\mathsf{P}(dx, dy) = \mathsf{P}_X(dx)\,\mathsf{P}_{Y|X}(x, dy)$.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Пусть $\mathsf{P} \in \mathcal{M}_+^1(Z)$.

Будем использовать следующий факт.

Если $(Y, \mathcal{S}_Y)$ является стандартным измеримым пространством, то в силу теоремы о дезинтеграции имеет место представление $\mathsf{P}(dx, dy) = \mathsf{P}_X(dx)\,\mathsf{P}_{Y|X}(x, dy)$.

Последовательно рассмотрим два примера, показывающие существование байесовского предиктора для задачи бинарной классификации и задачи обучения функций.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Утверждение 4.1 (случай бинарной классификации).

Пусть в качестве $(Y, \mathcal{S}_Y)$ выступает стандартное измеримое пространство $\left(\{0,1\}, 2^{\{0,1\}}\right)$, а в качестве функции потерь используется l_{01}.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Утверждение 4.1 (случай бинарной классификации).

Пусть в качестве $(Y, \mathcal{S}_Y)$ выступает стандартное измеримое пространство $\left(\{0,1\}, 2^{\{0,1\}}\right)$, а в качестве функции потерь используется l_{01}.

Тогда функция

$$g^*(x) := \begin{cases} 1, & \text{если} \quad \eta(x) := \mathsf{P}\left[\{Y=1\} \,|\, X=x\right] \geqslant \frac{1}{2}; \\ 0, & \text{иначе} \end{cases} \qquad (x \in X),$$

является байесовским предиктором.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Утверждение 4.1 (случай бинарной классификации).

Пусть в качестве $(Y, \mathcal{S}_Y)$ выступает стандартное измеримое пространство $\left(\{0,1\}, 2^{\{0,1\}}\right)$, а в качестве функции потерь используется l_{01}.

Тогда функция

$$g^*(x) := \begin{cases} 1, & \text{если} \quad \eta(x) := \mathsf{P}\left[\{Y=1\} \,|\, X=x\right] \geqslant \frac{1}{2}; \\ 0, & \text{иначе} \end{cases} \qquad (x \in X),$$

является байесовским предиктором.

◀ Для этого достаточно установить, что неравенство $R(g^*) \leqslant R(g)$ выполняется для любой измеримой функции $g \in \mathcal{S}_X \,|\, \mathcal{S}_Y$.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Действительно, с учётом теоремы о замене переменных в интеграле Лебега, получим

$$
\begin{aligned}
& R(g) \\
&= \int\limits_{X\times\{0,1\}} \mathbf{1}_{\{g(X)\neq Y\}} \mathsf{P}(dx,dy) \\
&= \int\limits_{X\times\{0,1\}} \mathbf{1}_{\{g(X)=0\}}\mathbf{1}_{\{Y=1\}} \mathsf{P}(dx,dy) + \int\limits_{X\times\{0,1\}} \mathbf{1}_{\{g(X)=1\}}\mathbf{1}_{\{Y=0\}} \mathsf{P}(dx,dy) \\
&= \int\limits_{\{x\in X\,:\,g(x)=0\}} \mathsf{P}\big[\{Y=1\}\,|\,X=x\big]\mathsf{P}_X(dx) + \int\limits_{\{x\in X\,:\,g(x)=1\}} \mathsf{P}\big[\{Y=0\}\,|\,X=x\big]\mathsf{P}_X(dx \\
&= \int\limits_{\{x\in X\,:\,g(x)=0\}} \eta(x)\mathsf{P}_X(dx) + \int\limits_{\{x\in X\,:\,g(x)=1\}} (1-\eta(x))\mathsf{P}_X(dx) \\
&= \int\limits_{X} \big[\mathbf{1}_{\{g(x)=0\}}\eta(x) + \mathbf{1}_{\{g(x)=1\}}(1-\eta(x))\big]\mathsf{P}_X(dx).
\end{aligned}
$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Обозначим

$$\varphi(g, x) := \mathbf{1}_{\{g(x)=0\}}\eta(x) + \mathbf{1}_{\{g(x)=1\}}(1 - \eta(x)) \quad (x \in X),$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Обозначим

$$\varphi(g, x) := \mathbf{1}_{\{g(x)=0\}}\eta(x) + \mathbf{1}_{\{g(x)=1\}}(1 - \eta(x)) \quad (x \in \mathsf{X}),$$

тогда

$$R(g) = \int_{\mathsf{X}} \varphi(g, x)\mathsf{P}_X(dx). \tag{1}$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Обозначим

$$\varphi(g, x) := \mathbf{1}_{\{g(x)=0\}}\eta(x) + \mathbf{1}_{\{g(x)=1\}}(1 - \eta(x)) \quad (x \in \mathsf{X}),$$

тогда

$$R(g) = \int_{\mathsf{X}} \varphi(g, x)\mathsf{P}_X(dx). \tag{1}$$

Заметим, что

$$g(x) = \begin{cases} 1, & \text{если} \quad \varphi(g, x) = 1 - \eta(x); \\ 0, & \text{если} \quad \varphi(g, x) = \eta(x) \end{cases} \qquad (x \in \mathsf{X}).$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Кроме того, по определению

$$g^*(x) = \begin{cases} 1, & \text{если} \quad \eta(x) \geqslant 1 - \eta(x); \\ 0, & \text{если} \quad \eta(x) < 1 - \eta(x) \end{cases} \qquad (x \in X).$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Кроме того, по определению

$$g^*(x) = \begin{cases} 1, & \text{если} \quad \eta(x) \geqslant 1 - \eta(x); \\ 0, & \text{если} \quad \eta(x) < 1 - \eta(x) \end{cases} \qquad (x \in X).$$

А значит

$$\varphi(g^*, x) \leqslant \varphi(g, x) \quad (x \in X). \tag{2}$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Кроме того, по определению

$$g^*(x) = \begin{cases} 1, & \text{если} \quad \eta(x) \geqslant 1 - \eta(x); \\ 0, & \text{если} \quad \eta(x) < 1 - \eta(x) \end{cases} \qquad (x \in X).$$

А значит

$$\varphi(g^*, x) \leqslant \varphi(g, x) \quad (x \in X). \tag{2}$$

Объединяя вместе (1) и (2), получим требуемое неравенство $R(g^*) \leqslant R(g)$.

■

Декомпозиция ошибки обучения

Существование байесовского предиктора

Утверждение 4.2 (случай обучения функций).

Пусть в качестве $(Y, \mathcal{S}_Y)$ выступает стандартное измеримое пространство $\big([0,1], \mathcal{B}([0,1])\big)$, а в качестве функции потерь используется l_{sq}.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Утверждение 4.2 (случай обучения функций).

Пусть в качестве $(Y, \mathcal{S}_Y)$ выступает стандартное измеримое пространство $\big([0,1], \mathcal{B}([0,1])\big)$, а в качестве функции потерь используется l_{sq}.

Тогда функция

$$g^*(x) := \mathbf{E}[Y \mid X = x] \quad (x \in X)$$

является байесовским предиктором.

Декомпозиция ошибки обучения

Существование байесовского предиктора

Утверждение 4.2 (случай обучения функций).

Пусть в качестве $(\mathsf{Y}, \mathcal{S}_{\mathsf{Y}})$ выступает стандартное измеримое пространство $\bigl([0,1], \mathcal{B}([0,1])\bigr)$, а в качестве функции потерь используется l_{sq}.

Тогда функция

$$g^*(x) \coloneqq \mathbf{E}[Y \mid X = x] \quad (x \in \mathsf{X})$$

является байесовским предиктором.

◀ Для этого будет достаточно установить справедливость следующего представления

$$R(g) = \|g - g^*\|^2_{\mathcal{L}^2(\mathsf{Z})} + R(g^*) \quad (g \in \mathcal{S}_{\mathsf{X}} \mid \mathcal{S}_{\mathsf{Y}}). \tag{3}$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Заметим, что $g^*(X) = \mathbf{E}[Y \mid X]$, а значит

$$\mathbf{E}[(g^*(X) - Y) \mid X] = g^*(X) - \mathbf{E}[Y \mid X] = g^*(X) - g^*(X) = 0 \quad \text{(P-п.н.)}. \quad (4)$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Заметим, что $g^*(X) = \mathbf{E}[Y \mid X]$, а значит

$$\mathbf{E}[(g^*(X) - Y) \mid X] = g^*(X) - \mathbf{E}[Y \mid X] = g^*(X) - g^*(X) = 0 \quad \text{(P-п.н.)}. \quad (4)$$

Зафиксируем произвольную функцию $g \in \mathcal{S}_X \mid \mathcal{S}_Y$ и запишем

$$\begin{aligned} & R(g) \\ & = \mathbf{E}[g(X) - Y]^2 = \mathbf{E}[g(X) - g^*(X) + g^*(X) - Y]^2 \\ & = \mathbf{E}[g(X) - g^*(X)]^2 + 2\mathbf{E}[(g(X) - g^*(X))(g^*(X) - Y)] + \mathbf{E}[g^*(X) - Y]^2 \\ & = \|g - g^*\|^2_{\mathcal{L}^2(Z)} + R(g^*) + 2\mathbf{E}[(g(X) - g^*(X))(g^*(X) - Y)]. \end{aligned} \quad (5)$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Используя свойства условного математического ожидания, получим

$$\begin{aligned}\mathbf{E}[(g(X)-g^*(X))(g^*(X)-Y)] &= \mathbf{E}\Big[\mathbf{E}[(g(X)-g^*(X))(g^*(X)-Y)\,|\,X]\Big] \\ &= \mathbf{E}\Big[(g(X)-g^*(X))\,\mathbf{E}[(g^*(X)-Y)\,|\,X]\Big] \\ &= \Big|\ \underset{(4)}{\text{Равенство}}\ \Big| = \mathbf{E}\Big[(g(X)-g^*(X))\cdot 0\Big] \\ &= 0.\end{aligned} \tag{6}$$

Декомпозиция ошибки обучения

Существование байесовского предиктора

Используя свойства условного математического ожидания, получим

$$\begin{aligned}\mathbf{E}[(g(X)-g^*(X))(g^*(X)-Y)] &= \mathbf{E}\big[\mathbf{E}[(g(X)-g^*(X))(g^*(X)-Y)\,|\,X]\big] \\ &= \mathbf{E}\big[(g(X)-g^*(X))\,\mathbf{E}[(g^*(X)-Y)\,|\,X]\big] \\ &= \Big|\ \underset{(4)}{\text{Равенство}}\ \Big| = \mathbf{E}\big[(g(X)-g^*(X))\cdot 0\big] \\ &= 0.\end{aligned} \tag{6}$$

Объединяя вместе (5) и (6), получим справедливость равенства (3).

■

Содержание

Декомпозиция ошибки обучения

Пример переобучения

Следующий пример является ответом на вопрос, почему в качестве класса гипотез не всегда разумно выбирать множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Декомпозиция ошибки обучения

Пример переобучения

Следующий пример является ответом на вопрос, почему в качестве класса гипотез не всегда разумно выбирать множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Пример 4.3.

В качестве множества объектов X возьмём единичный квадрат $[0,1]^2$ на плоскости, а в качестве множества меток Y возьмём $\{0,1\}$.

Декомпозиция ошибки обучения

Пример переобучения

Следующий пример является ответом на вопрос, почему в качестве класса гипотез не всегда разумно выбирать множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Пример 4.3.

В качестве множества объектов X возьмём единичный квадрат $[0,1]^2$ на плоскости, а в качестве множества меток Y возьмём $\{0,1\}$. Будем использовать функцию потерь l_{01}.

Декомпозиция ошибки обучения

Пример переобучения

Следующий пример является ответом на вопрос, почему в качестве класса гипотез не всегда разумно выбирать множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Пример 4.3.

В качестве множества объектов X возьмём единичный квадрат $[0,1]^2$ на плоскости, а в качестве множества меток Y возьмём $\{0,1\}$. Будем использовать функцию потерь l_{01}.

Внутри множества X зафиксируем произвольный квадрат R, имеющий площадь $\frac{1}{2}$, и стороны которого параллельны координатным осям.

Декомпозиция ошибки обучения

Пример переобучения

Следующий пример является ответом на вопрос, почему в качестве класса гипотез не всегда разумно выбирать множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Пример 4.3.

В качестве множества объектов X возьмём единичный квадрат $[0,1]^2$ на плоскости, а в качестве множества меток Y возьмём $\{0,1\}$. Будем использовать функцию потерь l_{01}.

Внутри множества X зафиксируем произвольный квадрат R, имеющий площадь $\frac{1}{2}$, и стороны которого параллельны координатным осям. В роли целевой функции g будет выступать $\mathbf{1}_R$.

Декомпозиция ошибки обучения

Пример переобучения

Следующий пример является ответом на вопрос, почему в качестве класса гипотез не всегда разумно выбирать множество всех измеримых отображений $\mathcal{S}_X \,|\, \mathcal{S}_Y$.

Пример 4.3.

В качестве множества объектов X возьмём единичный квадрат $[0,1]^2$ на плоскости, а в качестве множества меток Y возьмём $\{0,1\}$. Будем использовать функцию потерь l_{01}.

Внутри множества X зафиксируем произвольный квадрат R, имеющий площадь $\frac{1}{2}$, и стороны которого параллельны координатным осям. В роли целевой функции g будет выступать $\mathbf{1}_R$.

Вероятностная мера P на Z определяется через маргинальное распределение P_X, в роли которого выступает равномерное распределение на X.

Декомпозиция ошибки обучения

Пример переобучения

Пример 4.3 (продолжение).

Для каждого набора примеров $\mathbf{z} \in Z^*$ и объекта $x \in X$ через $S_1(\mathbf{z}, x)$ обозначим количество примеров вида $(x, 1)$ в наборе $\mathbf{z}$, а через $S_0(\mathbf{z}, x)$ обозначим количество примеров вида $(x, 0)$ в том же наборе.

Декомпозиция ошибки обучения

Пример переобучения

Пример 4.3 (продолжение).

Для каждого набора примеров $\mathbf{z} \in Z^*$ и объекта $x \in X$ через $S_1(\mathbf{z}, x)$ обозначим количество примеров вида $(x, 1)$ в наборе $\mathbf{z}$, а через $S_0(\mathbf{z}, x)$ обозначим количество примеров вида $(x, 0)$ в том же наборе.

Алгоритм обучения $\mathbf{z} \longmapsto h_{\mathbf{z}}$ $(\mathbf{z} \in Z^*)$, где

$$h_{\mathbf{z}}(x) := \begin{cases} 1, & \text{если} \quad S_1(\mathbf{z}, x) > S_0(\mathbf{z}, x); \\ 0, & \text{иначе} \end{cases} \qquad (x \in X)$$

является методом минимизации эмпирического риска.

Декомпозиция ошибки обучения

Пример переобучения

Пример 4.3 (продолжение).

Для каждого набора примеров $\mathbf{z} \in Z^*$ и объекта $x \in X$ через $S_1(\mathbf{z}, x)$ обозначим количество примеров вида $(x, 1)$ в наборе $\mathbf{z}$, а через $S_0(\mathbf{z}, x)$ обозначим количество примеров вида $(x, 0)$ в том же наборе.

Алгоритм обучения $\mathbf{z} \longmapsto h_{\mathbf{z}}$ $(\mathbf{z} \in Z^*)$, где

$$h_{\mathbf{z}}(x) := \begin{cases} 1, & \text{если} \quad S_1(\mathbf{z}, x) > S_0(\mathbf{z}, x); \\ 0, & \text{иначе} \end{cases} \qquad (x \in X)$$

является методом минимизации эмпирического риска.

Кроме того, при $\mathbf{z} \in Z^n$ $(n \in \mathbb{N})$ выполняется условие $r(l; h_{\mathbf{z}}, \mathbf{z}) = 0$ (P^n-п.н.). В то же время ожидаемый риск всегда $R(\mathsf{P}_X, l_{01}; g, h_{\mathbf{z}}) = \frac{1}{2}$.

Декомпозиция ошибки обучения

Пример переобучения

Заметим, что если в этом примере в качестве класса гипотез взять множество, состоящее из характеристических функций всех прямоугольников, стороны которых параллельны координатным осям, то минимальный риск такого класса также будет равен нулю.

Декомпозиция ошибки обучения

Пример переобучения

Заметим, что если в этом примере в качестве класса гипотез взять множество, состоящее из характеристических функций всех прямоугольников, стороны которых параллельны координатным осям, то минимальный риск такого класса также будет равен нулю.

При этом, любой метод минимизации эмпирического риска будет эффективно решать рассматривамую задачу.

Декомпозиция ошибки обучения

Пример переобучения

Заметим, что если в этом примере в качестве класса гипотез взять множество, состоящее из характеристических функций всех прямоугольников, стороны которых параллельны координатным осям, то минимальный риск такого класса также будет равен нулю.

При этом, любой метод минимизации эмпирического риска будет эффективно решать рассматривамую задачу.

С ростом размера обучающей выборки n ожидаемый риск может быть сколь угодно близко приближен к нулю с большой P^n-вероятностью.

Содержание

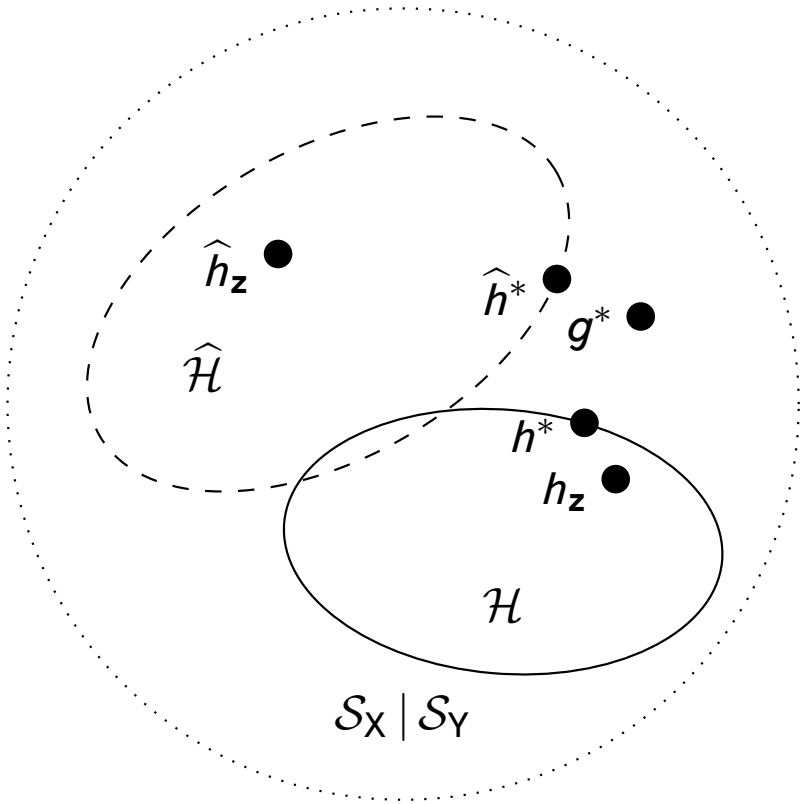
Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Обсудим общую схему поиска оптимального решения в задаче обучения, которая приведена на рисунке.

Она подразумевает выбор некоторого класса гипотез $\mathcal{H}$, минимальный риск которого приближает байесовскую ошибку.

Для этого класса подбирается алгоритм обучения. Например, в его роли может выступать некоторый метод минимизации эмпирического риска.



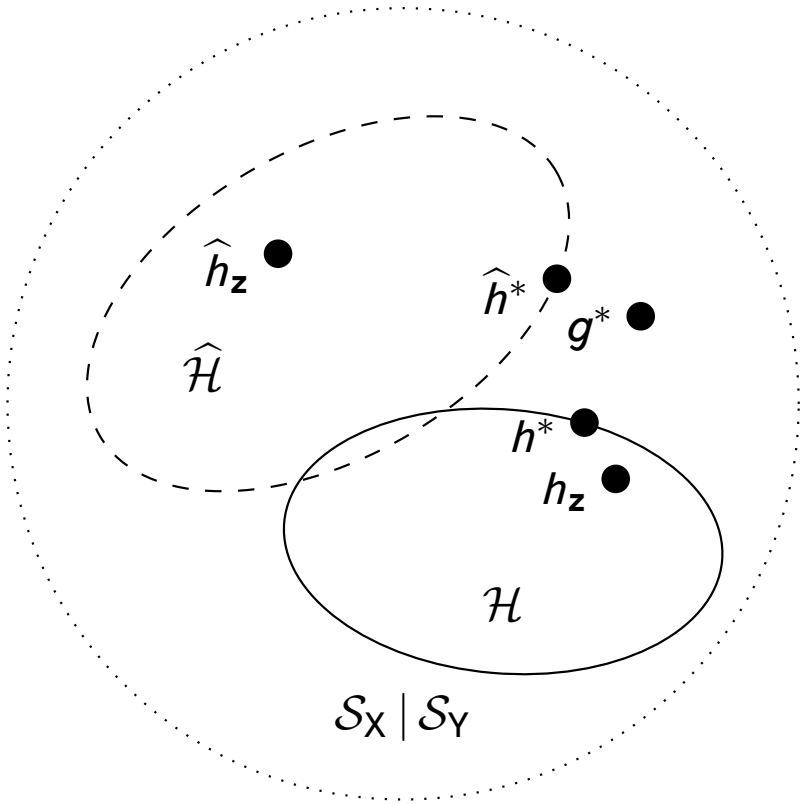
Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

С помощью этого алгоритма обучения по имеющейся обучающей выборке $\mathbf{z}$ строится гипотеза $h_{\mathbf{z}}$.

В качестве основной числовой характеристики, отражающей ошибку обучения, выступает избыточный риск $\mathcal{E}(h_{\mathbf{z}})$ этой гипотезы.

Избыточный риск отражает близость ожидаемого риска $R(h_{\mathbf{z}})$ к байесовской ошибке R^*.

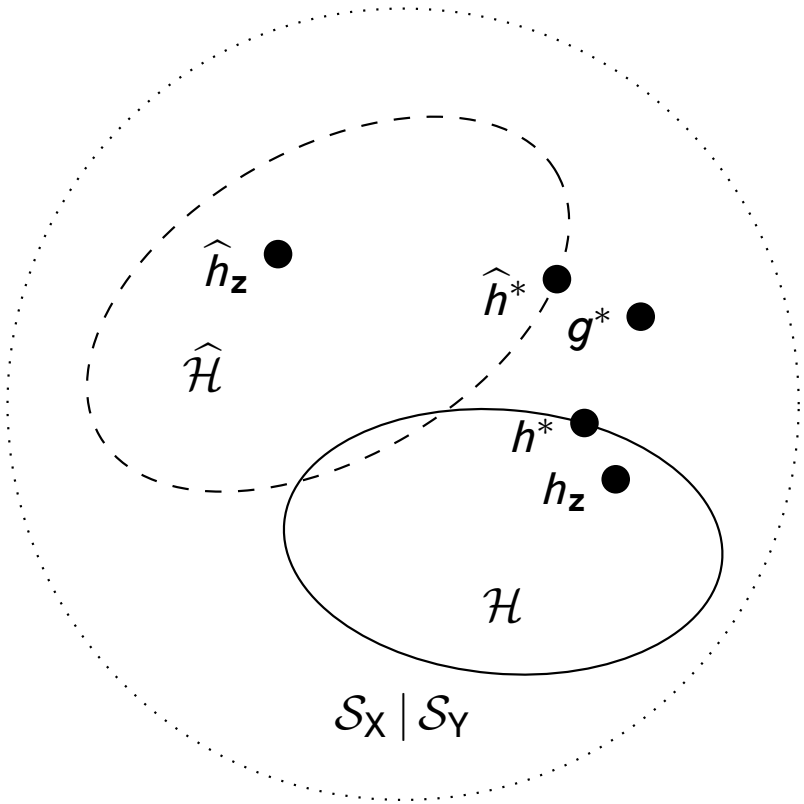


Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Это наилучшая точность, которая теоретически может быть достигнута.

В силу предположения о неизвестности вероятностной меры P, заданной на множестве примеров, избыточный риск не может быть точно вычислен. Он подлежит только приближённой оценке.

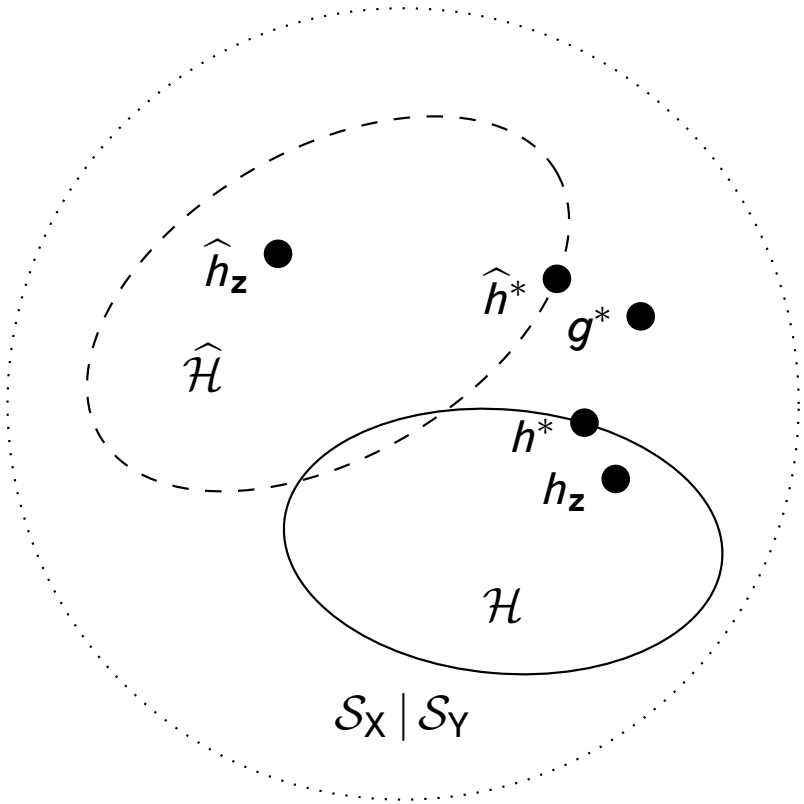


Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

В процессе поиска оптимального решения в задаче обучения может одновременно рассматриваться несколько альтернативных классов гипотез.

В этом случае говорят о задаче выбора модели.



Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Для любой гипотезы $h \in \mathcal{H}$ её избыточный риск может быть представлен в виде суммы двух неотрицательных слагаемых

$$\mathcal{E}(h) = \underbrace{R(h) - R(\mathcal{H})}_{\mathcal{E}_{\text{est}}} + \underbrace{R(\mathcal{H}) - R^*}_{\mathcal{E}_{\text{app}}}.$$

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Для любой гипотезы $h \in \mathcal{H}$ её избыточный риск может быть представлен в виде суммы двух неотрицательных слагаемых

$$\mathcal{E}(h) = \underbrace{R(h) - R(\mathcal{H})}_{\mathcal{E}_{\text{est}}} + \underbrace{R(\mathcal{H}) - R^*}_{\mathcal{E}_{\text{app}}}.$$

Величина $\mathcal{E}_{\text{est}}$ называется ошибкой оценивания, а величина $\mathcal{E}_{\text{app}}$ называется ошибкой аппроксимации (приближения).

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Для любой гипотезы $h \in \mathcal{H}$ её избыточный риск может быть представлен в виде суммы двух неотрицательных слагаемых

$$\mathcal{E}(h) = \underbrace{R(h) - R(\mathcal{H})}_{\mathcal{E}_{\text{est}}} + \underbrace{R(\mathcal{H}) - R^*}_{\mathcal{E}_{\text{app}}}.$$

Величина $\mathcal{E}_{\text{est}}$ называется ошибкой оценивания, а величина $\mathcal{E}_{\text{app}}$ называется ошибкой аппроксимации (приближения).

По аналогии с терминологией, принятой в математической статистике, ошибку оценивания также называют (индуктивным) смещением, а ошибку аппроксимации – отклонением.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Интуитивно понятно, что минимальный риск более «сложного» класса гипотез лучше приближает байесовскую ошибку.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Интуитивно понятно, что минимальный риск более «сложного» класса гипотез лучше приближает байесовскую ошибку.

Количество элементов конечного класса гипотез логично трактовать как его сложность.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Интуитивно понятно, что минимальный риск более «сложного» класса гипотез лучше приближает байесовскую ошибку.

Количество элементов конечного класса гипотез логично трактовать как его сложность.

При изучении модели агностического PAC-обучения будет показано, что с увеличением мощности конечного класса гипотез необходимо увеличивать и размер обучающей выборки для того, чтобы с заданной точностью и с заданной вероятностью ожидаемый риск гипотез из этого класса приближал его минимальный риск.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

В связи с этим разумной выглядит следующая стратегия.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

В связи с этим разумной выглядит следующая стратегия.

Вначале, увеличивая сложность, подбирается класс гипотез, позволяющий до приемлемого значения минимизировать ошибку аппроксимации $\mathcal{E}_{\text{app}}$.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

В связи с этим разумной выглядит следующая стратегия.

Вначале, увеличивая сложность, подбирается класс гипотез, позволяющий до приемлемого значения минимизировать ошибку аппроксимации $\mathcal{E}_{\text{app}}$.

Далее, за счет увеличения размера обучающей выборки при выбранном уровне доверия производится минимизация ошибки оценивания $\mathcal{E}_{\text{est}}$.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

В связи с этим разумной выглядит следующая стратегия.

Вначале, увеличивая сложность, подбирается класс гипотез, позволяющий до приемлемого значения минимизировать ошибку аппроксимации $\mathcal{E}_{\text{app}}$.

Далее, за счет увеличения размера обучающей выборки при выбранном уровне доверия производится минимизация ошибки оценивания $\mathcal{E}_{\text{est}}$.

К сожалению, эта стратегия не всегда является рабочей.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

В связи с этим разумной выглядит следующая стратегия.

Вначале, увеличивая сложность, подбирается класс гипотез, позволяющий до приемлемого значения минимизировать ошибку аппроксимации $\mathcal{E}_{\text{app}}$.

Далее, за счет увеличения размера обучающей выборки при выбранном уровне доверия производится минимизация ошибки оценивания $\mathcal{E}_{\text{est}}$.

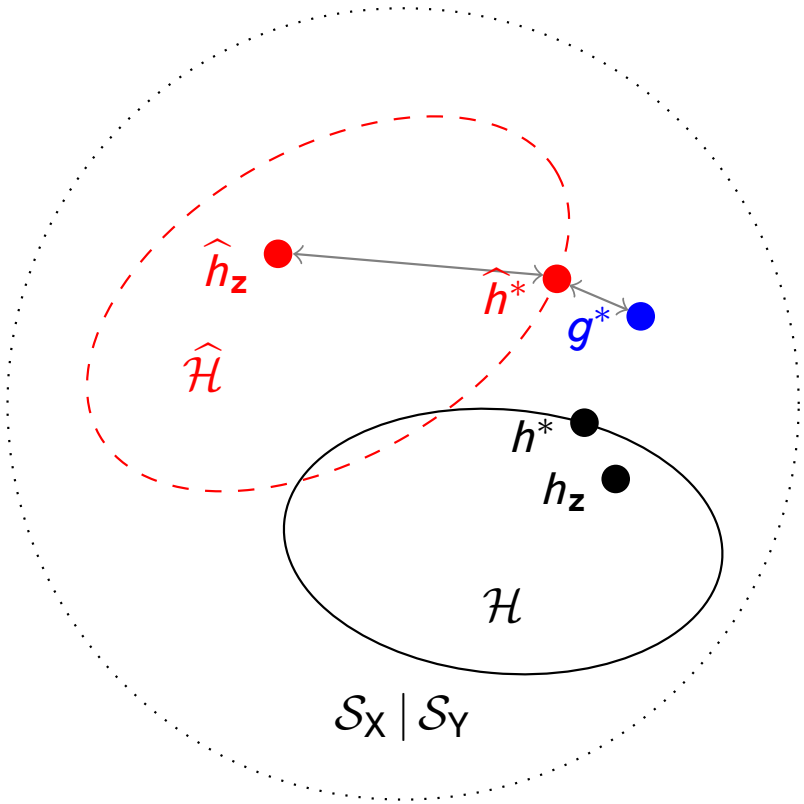
К сожалению, эта стратегия не всегда является рабочей.

Очень часто имеется фиксированная обучающая выборка и её размер за счёт добавления новых примеров не может быть увеличен.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

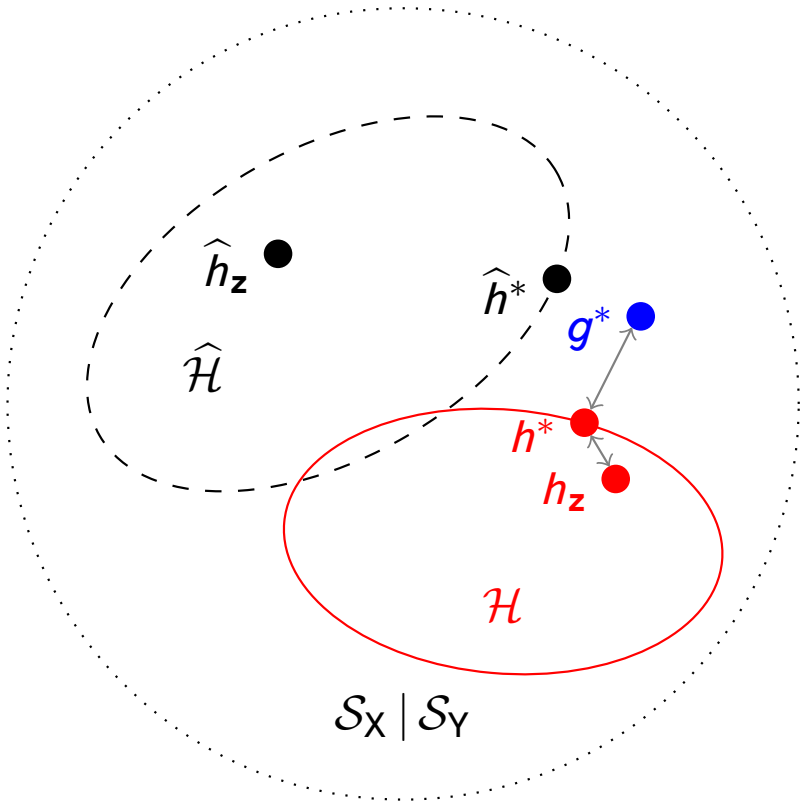
В этом случае приходится делать выбор либо между уменьшением ошибки аппроксимации и сопутствующим увеличением ошибки оценивания.



Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Либо сознательным увеличением ошибки аппроксимации, которое может привести к уменьшению ошибки оценивания.



Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Этот выбор обычно называют компромиссом между смещением и отклонением.

Декомпозиция ошибки обучения

Компромисс между ошибкой оценивания и приближения

Этот выбор обычно называют компромиссом между смещением и отклонением.

В дальнейшем, мы подробно займёмся вопросом минимизации ошибки оценивания.

Содержание

Содержание

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Прежде, чем непосредственно перейти к формулировке и обсуждению основного результата данного раздела, докажем два вспомогательных утверждения.

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Прежде, чем непосредственно перейти к формулировке и обсуждению основного результата данного раздела, докажем два вспомогательных утверждения.

Утверждение 4.1.

Пусть $P_X \in \mathcal{M}_+^1(X)$, $g, h \in \mathcal{S}_X \,|\, \mathcal{S}_Y$ и l – функция потерь. Предположим, что $(Y, \mathcal{S}_Y)$ – стандартное измеримое пространство.

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Прежде, чем непосредственно перейти к формулировке и обсуждению основного результата данного раздела, докажем два вспомогательных утверждения.

Утверждение 4.1.

Пусть $\mathsf{P}_X \in \mathcal{M}_+^1(\mathsf{X})$, $g, h \in \mathcal{S}_\mathsf{X} \,|\, \mathcal{S}_\mathsf{Y}$ и l – функция потерь. Предположим, что $(\mathsf{Y}, \mathcal{S}_\mathsf{Y})$ – стандартное измеримое пространство.

Тогда

$$R(\mathsf{P}_X, l; g, h) = \mathop{\mathbf{E}}_{x \sim \mathsf{P}_X} \big[l(h(x), g(x)) \big].$$

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Прежде, чем непосредственно перейти к формулировке и обсуждению основного результата данного раздела, докажем два вспомогательных утверждения.

Утверждение 4.1.

Пусть $\mathsf{P}_X \in \mathcal{M}_+^1(\mathsf{X})$, $g, h \in \mathcal{S}_\mathsf{X} \,|\, \mathcal{S}_\mathsf{Y}$ и l – функция потерь. Предположим, что $(\mathsf{Y}, \mathcal{S}_\mathsf{Y})$ – стандартное измеримое пространство.

Тогда

$$R(\mathsf{P}_X, l; g, h) = \mathop{\mathbf{E}}_{x \sim \mathsf{P}_X} \big[l(h(x), g(x)) \big].$$

◀ Обозначим через P вероятностную меру на Z, которая порождается маргинальным распределением P_X и целевой функцией g.

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Применяя теорему о дезинтеграции, получим

$$
\begin{aligned}
R(\mathsf{P}_X, l; g, h) &= \int_Z l(h(x), y)\mathsf{P}(dx, dy) \\
&= \int_X \mathsf{P}_X(dx) \int_Y l(h(x), y)\mathsf{P}_{Y|X}(x, dy) \\
&= \int_X \mathsf{P}_X(dx) \int_Y l(h(x), g(x))\mathsf{P}_{Y|X}(x, dy) \\
&= \int_X l(h(x), g(x))\mathsf{P}_X(dx).
\end{aligned}
$$

■

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Следующее утверждение позволяет уточнить понятие ожидаемого риска для реализуемого случая задачи обучения концептов.

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Следующее утверждение позволяет уточнить понятие ожидаемого риска для реализуемого случая задачи обучения концептов.

Утверждение 4.2.

Пусть $P_X \in \mathcal{M}_+^1(X)$ и $C', C \in \mathcal{S}_X$.

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Следующее утверждение позволяет уточнить понятие ожидаемого риска для реализуемого случая задачи обучения концептов.

Утверждение 4.2.

Пусть $\mathsf{P}_X \in \mathcal{M}_+^1(X)$ и $C', C \in \mathcal{S}_X$.

Тогда

$$R(\mathsf{P}_X; C', C) := R(\mathsf{P}_X, l_{01}; \mathbf{1}_{C'}, \mathbf{1}_C) = \mathsf{P}_X(C' \triangle C).$$

Отсутствие универсального алгоритма обучения

Вспомогательные утверждения

Следующее утверждение позволяет уточнить понятие ожидаемого риска для реализуемого случая задачи обучения концептов.

Утверждение 4.2.

Пусть $\mathsf{P}_X \in \mathcal{M}_+^1(\mathsf{X})$ и $C', C \in \mathcal{S}_\mathsf{X}$.

Тогда

$$R(\mathsf{P}_X; C', C) := R(\mathsf{P}_X, l_{01}; \mathbf{1}_{C'}, \mathbf{1}_C) = \mathsf{P}_X(C' \triangle C).$$

◀ Запишем

$$R(\mathsf{P}_X, l_{01}; \mathbf{1}_{C'}, \mathbf{1}_C) = \int_\mathsf{X} l_{01}(\mathbf{1}_{C'}, \mathbf{1}_C)\mathsf{P}_X(dx) = \int_\mathsf{X} \mathbf{1}_{C' \triangle C}\mathsf{P}_X(dx) = \mathsf{P}_X(C' \triangle C).$$

■

Содержание

Отсутствие универсального алгоритма обучения

Основной результат

Интуитивно под универсальным алгоритмом обучения можно понимать следующее.

Отсутствие универсального алгоритма обучения

Основной результат

Интуитивно под универсальным алгоритмом обучения можно понимать следующее.

Он потенциально применим к любой задаче обучения и формирует для неё приемлемое решение, ориентируясь только на те закономерности, которые можно извлечь из обучающей выборки.

Отсутствие универсального алгоритма обучения

Основной результат

Интуитивно под универсальным алгоритмом обучения можно понимать следующее.

Он потенциально применим к любой задаче обучения и формирует для неё приемлемое решение, ориентируясь только на те закономерности, которые можно извлечь из обучающей выборки.

Такой алгоритм не должен обладать предварительной информацией о структуре наиболее подходящего для решения конкретной задачи обучения класса гипотез, целевой функции или вероятностной меры, заданной на множестве примеров.

Отсутствие универсального алгоритма обучения

Основной результат

Рассмотренный ранее пример, демонстрирующий явление переобучения, подсказывает, что такого универсального алгоритма не существует.

Отсутствие универсального алгоритма обучения

Основной результат

Рассмотренный ранее пример, демонстрирующий явление переобучения, подсказывает, что такого универсального алгоритма не существует.

Следующая теорема позволяет формализовать эту догадку в виде строго математического утверждения.

Отсутствие универсального алгоритма обучения

Основной результат

Рассмотренный ранее пример, демонстрирующий явление переобучения, подсказывает, что такого универсального алгоритма не существует.

Следующая теорема позволяет формализовать эту догадку в виде строго математического утверждения.

Она имеет неформальное название no free lunch (не существует бесплатных обедов).

Отсутствие универсального алгоритма обучения

Основной результат

Теорема 4.1.

Пусть $n \in \mathbb{N}$ и $n < |X|/2$. Тогда для любого алгоритма обучения $\mathcal{A}$ существуют $C' \in 2^X$ и $P_X \in \mathcal{M}_+^1(X, 2^X)$ такие, что

$$P_X^n\left\{\mathbf{x} \in X^n \,:\, C = \mathcal{A}(\mathbf{1}_{C'} \circ \mathbf{x}),\, R(P_X; C', C) > \frac{1}{8}\right\} \geqslant \frac{1}{7}.$$

Отсутствие универсального алгоритма обучения

Основной результат

Теорема 4.1.

Пусть $n \in \mathbb{N}$ и $n < |\mathrm{X}|/2$. Тогда для любого алгоритма обучения $\mathcal{A}$ существуют $C' \in 2^{\mathrm{X}}$ и $\mathrm{P}_X \in \mathcal{M}_+^1\left(\mathrm{X}, 2^{\mathrm{X}}\right)$ такие, что

$$\mathrm{P}_X^n\left\{\mathbf{x} \in \mathrm{X}^n \,:\, C = \mathcal{A}(\mathbf{1}_{C'} \circ \mathbf{x}),\, R(\mathrm{P}_X; C', C) > \frac{1}{8}\right\} \geqslant \frac{1}{7}.$$

Прежде, чем перейти непосредственно к доказательству этой теоремы, докажем вспомогательное утверждение.

Отсутствие универсального алгоритма обучения

Основной результат

Утверждение 4.3.

Пусть $\widehat{\mathrm{X}} \subseteq \mathrm{X}$, $|\widehat{\mathrm{X}}| = 2n$ ($n \in \mathbb{N}$). Определим $\widehat{\mathrm{P}} \in \mathcal{M}_+^1\left(\widehat{\mathrm{X}}, 2^{\widehat{\mathrm{X}}}\right)$, полагая

$$\widehat{\mathrm{P}}\{x\} = \frac{1}{2n} \quad (x \in \hat{\mathrm{X}}).$$

Отсутствие универсального алгоритма обучения

Основной результат

Утверждение 4.3.

Пусть $\widehat{\mathsf{X}} \subseteq \mathsf{X}$, $|\widehat{\mathsf{X}}| = 2n$ ($n \in \mathbb{N}$). Определим $\widehat{\mathsf{P}} \in \mathcal{M}_+^1\big(\widehat{\mathsf{X}}, 2^{\widehat{\mathsf{X}}}\big)$, полагая

$$\widehat{\mathsf{P}}\{x\} = \frac{1}{2n} \quad (x \in \hat{\mathsf{X}}).$$

Тогда для любого алгоритма обучения $\widehat{\mathcal{A}}$ справедливо неравенство

$$\max_{C \in 2^{\widehat{\mathsf{X}}}} \mathop{\mathbf{E}}_{\mathbf{x} \sim \widehat{\mathsf{P}}^n} R\big(\widehat{\mathsf{P}}; C, \widehat{\mathcal{A}}(\mathbf{1}_C \circ \mathbf{x})\big) \geqslant \frac{1}{4}. \tag{7}$$

Отсутствие универсального алгоритма обучения

Основной результат

Утверждение 4.3.

Пусть $\widehat{\mathsf{X}} \subseteq \mathsf{X}$, $|\widehat{\mathsf{X}}| = 2n$ ($n \in \mathbb{N}$). Определим $\widehat{\mathsf{P}} \in \mathcal{M}_+^1\big(\widehat{\mathsf{X}}, 2^{\widehat{\mathsf{X}}}\big)$, полагая

$$\widehat{\mathsf{P}}\{x\} = \frac{1}{2n} \quad (x \in \hat{\mathsf{X}}).$$

Тогда для любого алгоритма обучения $\widehat{\mathcal{A}}$ справедливо неравенство

$$\max_{C \in 2^{\widehat{\mathsf{X}}}} \mathop{\mathbf{E}}_{\mathbf{x} \sim \widehat{\mathsf{P}}^n} R\big(\widehat{\mathsf{P}}; C, \widehat{\mathcal{A}}(\mathbf{1}_C \circ \mathbf{x})\big) \geqslant \frac{1}{4}. \tag{7}$$

◀ Обозначим $N = 2^{2n}$ и $M := (2n)^n$. Занумеруем $C_1, C_2, \ldots, C_N$ все концепты из $2^{\widehat{\mathsf{X}}}$.

Отсутствие универсального алгоритма обучения

Основной результат

Обозначим

$$R_i(\mathbf{x}) := R\big(\widehat{\mathrm{P}}; C_i, \widehat{\mathcal{A}}(\mathbf{1}_{C_i} \circ \mathbf{x})\big) \qquad (\mathbf{x} \in \widehat{\mathrm{X}}^n;\ i = 1, \ldots, N). \tag{8}$$

Отсутствие универсального алгоритма обучения

Основной результат

Обозначим

$$R_i(\mathbf{x}) := R\big(\widehat{\mathrm{P}}; C_i, \widehat{\mathcal{A}}(\mathbf{1}_{C_i} \circ \mathbf{x})\big) \qquad (\mathbf{x} \in \widehat{\mathrm{X}}^n;\ i = 1, \ldots, N). \tag{8}$$

Заметим, что

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \widehat{\mathrm{P}}^n} R_i(\mathbf{x}) = \frac{1}{M} \sum_{\mathbf{x} \in \widehat{\mathrm{X}}^n} R_i(\mathbf{x}) \qquad (i = 1, \ldots, N).$$

Отсутствие универсального алгоритма обучения

Основной результат

Обозначим

$$R_i(\mathbf{x}) := R\big(\widehat{\mathrm{P}}; C_i, \widehat{\mathcal{A}}(\mathbf{1}_{C_i} \circ \mathbf{x})\big) \qquad (\mathbf{x} \in \widehat{\mathrm{X}}^n;\ i = 1, \ldots, N). \tag{8}$$

Заметим, что

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \widehat{\mathrm{P}}^n} R_i(\mathbf{x}) = \frac{1}{M} \sum_{\mathbf{x} \in \widehat{\mathrm{X}}^n} R_i(\mathbf{x}) \qquad (i = 1, \ldots, N).$$

У любой конечной последовательности неотрицательных чисел максимальное значение не меньше, чем среднее арифметическое значение, которое в свою очередь не меньше минимального значения.

Отсутствие универсального алгоритма обучения

Основной результат

Следовательно,

$$\max_{1\leqslant i\leqslant N} \mathop{\mathbf{E}}_{\mathbf{x}\sim\widehat{\mathsf{P}}^n} R_i(\mathbf{x}) \geqslant \frac{1}{N}\sum_{i=1}^{N} \mathop{\mathbf{E}}_{\mathbf{x}\sim\widehat{\mathsf{P}}^n} R_i(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{\mathbf{x}\in\widehat{\mathsf{X}}^n} R_i(\mathbf{x}) = \\ = \frac{1}{M}\sum_{\mathbf{x}\in\widehat{\mathsf{X}}^n}\frac{1}{N}\sum_{i=1}^{N} R_i(\mathbf{x}) \geqslant \min_{\mathbf{x}\in\widehat{\mathsf{X}}^n}\frac{1}{N}\sum_{i=1}^{N} R_i(\mathbf{x}). \tag{9}$$

Зафиксируем произвольный набор $\mathbf{x}\in\widehat{\mathsf{X}}^n$.

Отсутствие универсального алгоритма обучения

Основной результат

Следовательно,

$$\max_{1\leqslant i\leqslant N} \mathop{\mathbf{E}}_{\mathbf{x}\sim\widehat{\mathsf{P}}^n} R_i(\mathbf{x}) \geqslant \frac{1}{N}\sum_{i=1}^{N} \mathop{\mathbf{E}}_{\mathbf{x}\sim\widehat{\mathsf{P}}^n} R_i(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{\mathbf{x}\in\widehat{\mathsf{X}}^n} R_i(\mathbf{x}) = \\ = \frac{1}{M}\sum_{\mathbf{x}\in\widehat{\mathsf{X}}^n}\frac{1}{N}\sum_{i=1}^{N} R_i(\mathbf{x}) \geqslant \min_{\mathbf{x}\in\widehat{\mathsf{X}}^n}\frac{1}{N}\sum_{i=1}^{N} R_i(\mathbf{x}). \tag{9}$$

Зафиксируем произвольный набор $\mathbf{x} \in \widehat{\mathsf{X}}^n$.

Введём подмножество $N_{\mathbf{x}} \subset \widehat{\mathsf{X}}$, состоящее из всех объектов, не встречающихся в наборе $\mathbf{x}$.

Отсутствие универсального алгоритма обучения

Основной результат

Следовательно,

$$\max_{1\leqslant i\leqslant N} \mathop{\mathbf{E}}_{\mathbf{x}\sim\widehat{\mathsf{P}}^n} R_i(\mathbf{x}) \geqslant \frac{1}{N}\sum_{i=1}^{N} \mathop{\mathbf{E}}_{\mathbf{x}\sim\widehat{\mathsf{P}}^n} R_i(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{\mathbf{x}\in\widehat{\mathsf{X}}^n} R_i(\mathbf{x}) = \\ = \frac{1}{M}\sum_{\mathbf{x}\in\widehat{\mathsf{X}}^n}\frac{1}{N}\sum_{i=1}^{N} R_i(\mathbf{x}) \geqslant \min_{\mathbf{x}\in\widehat{\mathsf{X}}^n}\frac{1}{N}\sum_{i=1}^{N} R_i(\mathbf{x}). \tag{9}$$

Зафиксируем произвольный набор $\mathbf{x} \in \widehat{\mathsf{X}}^n$.

Введём подмножество $N_{\mathbf{x}} \subset \widehat{\mathsf{X}}$, состоящее из всех объектов, не встречающихся в наборе $\mathbf{x}$.

Очевидно, что $|N_{\mathbf{x}}| \geqslant n$.

Отсутствие универсального алгоритма обучения

Основной результат

Обозначим

$$\widehat{C}_i := \widehat{\mathcal{A}}(\mathbf{1}_{C_i} \circ \mathbf{x}) \quad (i = 1, \ldots, N),$$

Отсутствие универсального алгоритма обучения

Основной результат

Обозначим

$$\widehat{C}_i := \widehat{\mathcal{A}}(\mathbf{1}_{C_i} \circ \mathbf{x}) \quad (i = 1, \ldots, N),$$

тогда

$$R_i(\mathbf{x}) = \widehat{\mathsf{P}}(C_i \triangle \widehat{C}_i) = \frac{1}{2n} \sum_{x \in \widehat{\mathsf{X}}} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x) \geqslant \frac{1}{2|N_{\mathbf{x}}|} \sum_{x \in N_{\mathbf{x}}} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x).$$

Отсутствие универсального алгоритма обучения

Основной результат

Обозначим

$$\widehat{C}_i := \widehat{\mathcal{A}}(\mathbf{1}_{C_i} \circ \mathbf{x}) \quad (i = 1, \ldots, N),$$

тогда

$$R_i(\mathbf{x}) = \widehat{\mathsf{P}}(C_i \triangle \widehat{C}_i) = \frac{1}{2n} \sum_{x \in \widehat{\mathsf{X}}} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x) \geqslant \frac{1}{2|N_{\mathbf{x}}|} \sum_{x \in N_{\mathbf{x}}} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x).$$

Следовательно,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} R_i(\mathbf{x}) &\geqslant \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2|N_{\mathbf{x}}|} \sum_{x \in N_{\mathbf{x}}} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x) \\ &= \frac{1}{2|N_{\mathbf{x}}|} \sum_{x \in N_{\mathbf{x}}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x) \geqslant \frac{1}{2} \min_{x \in N_{\mathbf{x}}} \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x). \end{aligned} \tag{10}$$

Отсутствие универсального алгоритма обучения

Основной результат

Зафиксируем произвольный $x \in N_{\mathbf{x}}$.

Отсутствие универсального алгоритма обучения

Основной результат

Зафиксируем произвольный $x \in N_{\mathbf{x}}$.

Множество всех концептов разобьём на $\frac{N}{2}$ различных пар вида (C_{i_1}, C_{i_2}) таких, что $i_1 < i_2$ и $C_{i_1} \triangle C_{i_2} = \{x\}$.

Отсутствие универсального алгоритма обучения

Основной результат

Зафиксируем произвольный $x \in N_{\mathbf{x}}$.

Множество всех концептов разобьём на $\frac{N}{2}$ различных пар вида (C_{i_1}, C_{i_2}) таких, что $i_1 < i_2$ и $C_{i_1} \triangle C_{i_2} = \{x\}$.

Учитывая равенство $\widehat{C}_{i_1} = \widehat{C}_{i_2}$, получим

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x) = \frac{1}{N} \sum_{\substack{i_1 < i_2 \\ C_{i_1} \triangle C_{i_2} = \{x\}}} \left[\mathbf{1}_{C_{i_1} \triangle \widehat{C}_{i_1}}(x) + \mathbf{1}_{C_{i_2} \triangle \widehat{C}_{i_1}}(x)\right] = \frac{1}{N} \cdot \frac{N}{2} = \frac{1}{2}. \tag{11}$$

Отсутствие универсального алгоритма обучения

Основной результат

Зафиксируем произвольный $x \in N_{\mathbf{x}}$.

Множество всех концептов разобьём на $\frac{N}{2}$ различных пар вида (C_{i_1}, C_{i_2}) таких, что $i_1 < i_2$ и $C_{i_1} \triangle C_{i_2} = \{x\}$.

Учитывая равенство $\widehat{C}_{i_1} = \widehat{C}_{i_2}$, получим

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{C_i \triangle \widehat{C}_i}(x) = \frac{1}{N} \sum_{\substack{i_1 < i_2 \\ C_{i_1} \triangle C_{i_2} = \{x\}}} \left[\mathbf{1}_{C_{i_1} \triangle \widehat{C}_{i_1}}(x) + \mathbf{1}_{C_{i_2} \triangle \widehat{C}_{i_1}}(x)\right] = \frac{1}{N} \cdot \frac{N}{2} = \frac{1}{2}. \tag{11}$$

Объединяя вместе (8), (9), (10) и (11), получим требуемое равенство (7).

■

Отсутствие универсального алгоритма обучения

Основной результат

◀ (доказательство теоремы 4.1) По условию существует подмножество $\widehat{X} \subseteq X$ такое, что $|\widehat{X}| = 2n$.

Отсутствие универсального алгоритма обучения

Основной результат

◀ (доказательство теоремы 4.1) По условию существует подмножество $\widehat{X} \subseteq X$ такое, что $|\widehat{X}| = 2n$.

Определим $P_X \in \mathcal{M}^1_+(X, 2^X)$, полагая

$$P_X(A) := \widehat{P}(A \cap \widehat{X}) \qquad (A \in 2^X),$$

где $\widehat{P}$ – вероятностная мера из формулировки утв. 4.3.

Отсутствие универсального алгоритма обучения

Основной результат

◀ (доказательство теоремы 4.1) По условию существует подмножество $\widehat{X} \subseteq X$ такое, что $|\widehat{X}| = 2n$.

Определим $P_X \in \mathcal{M}^1_+(X, 2^X)$, полагая

$$P_X(A) := \widehat{P}(A \cap \widehat{X}) \qquad (A \in 2^X),$$

где $\widehat{P}$ – вероятностная мера из формулировки утв. 4.3.

Определим алгоритм обучения

$$\widehat{\mathcal{A}}(\mathbf{z}) := \mathcal{A}(\mathbf{z}) \cap \widehat{X} \qquad (\mathbf{z} \in Z^*),$$

Отсутствие универсального алгоритма обучения

Основной результат

◀ (доказательство теоремы 4.1) По условию существует подмножество $\widehat{\mathsf{X}} \subseteq \mathsf{X}$ такое, что $|\widehat{\mathsf{X}}| = 2n$.

Определим $\mathsf{P}_X \in \mathcal{M}_+^1\left(\mathsf{X}, 2^{\mathsf{X}}\right)$, полагая

$$\mathsf{P}_X(A) := \widehat{\mathsf{P}}(A \cap \widehat{\mathsf{X}}) \qquad \left(A \in 2^{\mathsf{X}}\right),$$

где $\widehat{\mathsf{P}}$ – вероятностная мера из формулировки утв. 4.3.

Определим алгоритм обучения

$$\widehat{\mathcal{A}}(\mathbf{z}) := \mathcal{A}(\mathbf{z}) \cap \widehat{\mathsf{X}} \qquad (\mathbf{z} \in \mathsf{Z}^*),$$

тогда

$$R\left(\mathsf{P}_X; C, \mathcal{A}(\mathbf{1}_C \circ \mathbf{x})\right) = R\left(\widehat{\mathsf{P}}; C, \widehat{\mathcal{A}}(\mathbf{1}_C \circ \mathbf{x})\right) \qquad \left(C \in 2^{\widehat{\mathsf{X}}}, \mathbf{x} \in \widehat{\mathsf{X}}^n\right).$$

Отсутствие универсального алгоритма обучения

Основной результат

Из утв. 4.3 следует существование концепта $C' \in 2^{\widehat{X}} \subseteq 2^X$, для которого

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathrm{P}_X^n} R\big(\mathrm{P}_X; C', \mathcal{A}(\mathbf{1}_{C'} \circ \mathbf{x})\big) \geqslant \mathop{\mathbf{E}}_{\mathbf{x} \sim \widehat{\mathrm{P}}^n} R\big(\widehat{\mathrm{P}}; C', \widehat{\mathcal{A}}(\mathbf{1}_{C'} \circ \mathbf{x})\big) \geqslant \frac{1}{4}.$$

Отсутствие универсального алгоритма обучения

Основной результат

Из утв. 4.3 следует существование концепта $C' \in 2^{\widehat{X}} \subseteq 2^X$, для которого

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim \mathsf{P}_X^n} R\big(\mathsf{P}_X; C', \mathcal{A}(\mathbf{1}_{C'} \circ \mathbf{x})\big) \geqslant \mathop{\mathbf{E}}_{\mathbf{x}\sim \widehat{\mathsf{P}}^n} R\big(\widehat{\mathsf{P}}; C', \widehat{\mathcal{A}}(\mathbf{1}_{C'} \circ \mathbf{x})\big) \geqslant \frac{1}{4}.$$

Используя следствие 2.1 из неравенства Маркова, окончательно получим

$$\mathsf{P}_X^n\left\{\mathbf{x} \in X^n \,:\, R\big(\mathsf{P}_X; C', \mathcal{A}(\mathbf{1}_{C'} \circ \mathbf{x})\big) > \frac{1}{8}\right\} \geqslant \frac{8}{7} \cdot \left(\frac{1}{4} + \frac{7}{8} - 1\right) = \frac{1}{7}.$$

■