Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

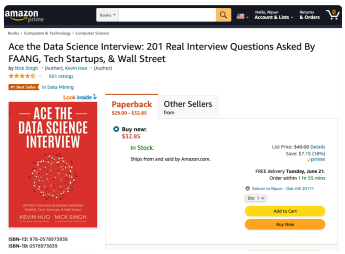
Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

30 вопросов на собеседовании по машинному обучению и учебное пособие по машинному обучению

КОНСУЛЬТАЦИИ ПО КАРЬЕРЕ В ОБЛАСТИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Инженеры по машинному обучению и специалисты по данным, занимающиеся машинным обучением, найдут это учебное пособие по наиболее важным темам, затронутым во время интервью по машинному обучению, полезным для их следующего поиска работы. Если вы аналитик данных, это руководство, вероятно, излишне... вам лучше попрактиковаться в SQL-вопросах для интервью . Но для остальных из вас, ярых поклонников глубокого обучения, пожалуйста, воспользуйтесь этими **30 вопросами для интервью по машинному обучению** , которые недавно задали такие компании, как Google, Netflix и Stripe.

Кевин Хо, бывший сотрудник Facebook Data Scientist, который сейчас работает в хедж-фонде, и я подробно решили 8 задач. Мы оставили решение остальных проблем в нашей книге « Ace the Data Science Interview» (теперь она доступна на Amazon!).



Кроме того, некоторые вопросы по машинному обучению из этой статьи и нашей книги можно бесплатно найти на DataLemur (где размещены сотни вопросов по науке о данных и интервью по машинному обучению).

Это руководство охватывает:

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

- Часто задаваемые темы машинного обучения для рассмотрения на технических собеседованиях . Мы кратко рассмотрим математику, стоящую за ML , линейной регрессией , уменьшением размерности , кластеризацией и многим другим.
- 30 задач на собеседовании по машинному обучению
- 8 ответов на вопросы интервью по ML
- Как получить больше вопросов для интервью по науке о данных

Часто задаваемые темы машинного обучения во время технических интервью

Вот общий обзор областей, затронутых в интервью ведущих технологических компаний и Уолл-Стрит. Чтобы глубже изучить эти темы, ознакомьтесь с некоторыми из рекомендуемых нами учебников, онлайн-курсов и учебных курсов, которые мы представили в нашем Руководстве по науке о данных . Также может быть полезно прочитать наше руководство «40 вопросов для интервью по науке о вероятностях и статистических данных» , поскольку эта математика лежит в основе многих методов машинного обучения, описанных ниже.

Математические предпосылки

Случайные переменные

Случайные величины являются основной темой теории вероятностей и статистики, и интервьюеры, как правило, ищут понимания принципов и базовой способности ими манипулировать. Случайная величина — это величина с соответствующим распределением вероятностей, которая может быть либо дискретной (исчисляемый диапазон), либо непрерывной (несчетный диапазон). Для дискретной случайной величины существует функция массы вероятности, а для непрерывной случайной величины — функция плотности вероятности.

$$\text{Discrete: } \sum_{x \in X} p(x) = 1, \text{ Continuous: } \int_{-\infty}^{\infty} p(x)dx = 1$$

Кумулятивная функция распределения определяется как:

$$F(x) = p(X \leq x)$$

Для любой заданной случайной величины X она обладает следующими свойствами (ниже мы предполагаем, что X непрерывна, но аналогично верно для дискретных случайных величин). Ожидание (среднее значение) определяется по формуле:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$

а дисперсия определяется как:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X^2]$$

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google. Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ
ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ
ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL
14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ
ОБО МНЕ

Для любых заданных случайных величин X и Y ковариация, линейная мера отношения, определяется следующим образом:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

а нормализация ковариации - это корреляция между X и y:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Распределения вероятностей

Есть много вероятностных распределений, и интервьюеры, как правило, не проверяют, запомнили ли вы конкретные свойства каждого из них (хотя знать основы полезно), а в большей степени, чтобы вы могли правильно применить их к конкретным ситуациям. Из-за этого наиболее часто обсуждаемым в интервью по науке о данных является нормальное распределение, которое имеет множество реальных применений. Для одной переменной плотность вероятности определяется следующим образом для параметра среднего значения и дисперсии:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp - \left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Для подгонки параметров существует два основных метода. Целью оценки максимального правдоподобия (MLE) является оценка наиболее вероятных параметров с учетом функции правдоподобия:

$$\theta_{MLE} = \text{argmax}_\theta L(\theta), \text{ where } L(\theta) = p(x_1, \ldots x_n|\theta)$$

Поскольку предполагается, что значения X равны iid, функция правдоподобия принимает вид:

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Удобно брать логарифмы (поскольку логарифм — монотонно возрастающая функция, максимизация логарифмического правдоподобия эквивалентна максимизации правдоподобия):

$$\log L(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

Другой способ подбора параметров — максимальная апостериорная оценка (MAP), которая предполагает априорное распределение.

$$\theta_{MAP} = \text{argmax}_\theta p(\theta)p(x_1 \ldots x_n|\theta)$$

где применяется аналогичное логарифмическое правдоподобие из предыдущего.

В то время как о нормальном распределении часто спрашивают в интервью, важно знать подробности и варианты использования других известных распределений, таких как равномерное, пуассоновское, биномиальное и геометрическое распределение.

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

Никогда не помешает иметь возможность делать выводы для ожидания, дисперсии или других более высоких моментов.

Линейная алгебра

Как правило, интервьюеры не ожидают, что вы будете глубоко вникать в линейную алгебру, если не будете делать особый акцент на машинном обучении. Тем не менее, обзор основ все же полезен, поскольку он помогает понять различные алгоритмы и теоретические основы. В линейной алгебре есть много подтем, но одна подтема, которую стоит кратко обсудить, — это собственные значения и собственные векторы. Механически для некоторой квадратной матрицы A вектор x является собственным вектором матрицы A, если:

$$Ax = \lambda x$$

Поскольку матрица представляет собой линейное преобразование, собственные векторы — это случаи, когда результирующее преобразование матрицы на этом векторе приводит к тому же направлению, что и раньше, хотя и с некоторым коэффициентом масштабирования (собственными значениями). Существует много реальных случаев использования собственных значений и собственных векторов: например, определение ориентации больших наборов данных (обсуждается в PCA) или для динамических систем (как система колеблется и как быстро она стабилизируется).

Разложение квадратной матрицы на собственные векторы называется собственным разложением. Обратите внимание, что хотя не все матрицы квадратные, с помощью разложения по сингулярным значениям (SVD) каждая матрица имеет разложение:

$$A = U\Sigma V^T$$

Хотя математические детали выходят за рамки этого обсуждения, как собственное разложение, так и SVD заслуживают подробного изучения перед вашим техническим собеседованием.

Компромисс смещения и дисперсии

Эту тему иногда задают в интервью из-за ее актуальности с переоснащением и выбором модели. С любой моделью мы обычно пытаемся оценить истинную основу:

$$y = f(x) + w$$

через данные, где w обычно представляет собой шум с нулевым средним значением и гауссовой случайной величиной. Как упоминалось ранее, MLE и MAP являются разумными способами вывода параметров. Чтобы оценить, насколько хорошо подходит модель, мы можем разложить ошибку y следующим образом:

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

1) смещение (насколько хорошо значения приближаются к истинным базовым значениям f(x))

2) дисперсия (насколько прогноз изменяется на основе данных обучения)

3) неустранимая ошибка (из-за зашумленности процессов наблюдения.

Существует компромисс между предвзятостью и дисперсией, и это полезная основа для размышлений о том, как работают разные модели. Общая цель состоит в том, чтобы контролировать переоснащение (а не обобщение выборки) для создания стабильных и точных моделей.

Линейная регрессия

Этот метод является одним из наиболее часто изучаемых методов и имеет множество применений в реальной жизни, начиная от прогнозирования цен на жилье и заканчивая изучением эффективности медицинских испытаний. Интервьюеры, спрашивающие об этом, как правило, пытаются оценить ваше понимание основных формулировок, а иногда и актуальность знания некоторых теорий для применения в реальной жизни.

В линейной регрессии цель состоит в том, чтобы оценить y = f (x) в следующей форме:

$$y = X\beta$$

где X — матрица точек данных, а β — вектор весов. В контексте метода наименьших квадратов линейная регрессия минимизирует остаточную сумму квадратов (RSS), которая определяется как:

$$RSS(\beta) = (y - X\beta)^T(y - X\beta)$$

В регрессии можно использовать MLE для оценки значений β с помощью многомерного Гаусса:

$$Y \sim N(X\beta, \sigma^2 I)$$

что приводит к результатам, аналогичным минимизации RSS. Для контекста MAP могут быть априорные значения для β, что приводит к регрессии хребта, которая штрафует веса для предотвращения переобучения. В регрессии Риджа целевая функция становится минимизирующей:

$$(y - X\beta)^T(y - X\beta) + \lambda|\beta|_2^2$$

Уменьшение размерности

Анализ основных компонентов

Эта тема реже встречается в интервью, но часто упоминается при обсуждении предварительной обработки данных или разработки функций. Разложение данных на меньший набор переменных очень полезно

для обобщения и визуализации данных. Этот общий процесс называется уменьшением размерности. Одним из распространенных методов уменьшения размерности является анализ основных компонентов (PCA), который реконструирует данные в более низком размере. Он ищет небольшое количество линейных комбинаций вектора x (скажем, p-мерного), чтобы объяснить дисперсию внутри x. В частности, мы хотим найти вектор w весов, чтобы мы могли определить следующую линейную комбинацию:

$$y_i = w_i^T x = \sum_{j=1}^{p} w_{ij} x_j$$

с учетом следующего:

$$y_i \text{ is uncorrelated with } y_j, var(y_i) \text{ is maximized}$$

Следовательно, у нас есть следующее процедурное описание, в котором сначала мы находим первую компоненту с максимальной дисперсией, а затем вторую, некоррелированную с первой, и продолжаем эту процедуру итеративно. Идея состоит в том, чтобы закончить, скажем, k измерением таким, что

$$y_1, \dots y_k \text{ explain the majority of variance, } k << p$$

Используя некоторую алгебру, окончательный результат представляет собой собственное разложение ковариационной матрицы X, в результате чего первый главный компонент является собственным вектором, соответствующим наибольшему собственному значению, и так далее.

Классификация

Общая структура

Классификация обычно задается во время интервью из-за обилия реальных приложений. Технологические компании любят спрашивать о классификации клиентов и пользователей по разным сегментам.

Цель классификации состоит в том, чтобы отнести заданную точку данных к одному из K классов вместо непрерывного значения (как в регрессии). Существует два типа моделей. Первый является генеративным, который моделирует совместное распределение вероятностей между X и Y. То есть для входных данных X мы хотим классифицировать произвольную точку данных x со следующей меткой класса:

$$\hat{y} = \backslash\text{argmax}_k p(x, Y = k)$$

Это совместное распределение между X и Y определяется как:

$$p(X, Y) = p(Y|X)p(X)$$

и для каждого заданного класса k имеем:

$$p_k(X) = p(X|k)p(k)$$

так что:

$$\hat{y} = \text{argmax}_k p(Y = k|x)$$

Результат максимизации апостериорных значений означает, что между классами будут границы решений, при которых результирующая апостериорная вероятность равна.

Второй - дискриминативный, который напрямую изучает границу решения, выбирая класс, который максимизирует апостериорное распределение вероятностей:

$$\hat{y} = \text{argmax}_k p(Y = k|x)$$

Таким образом, оба метода в конечном итоге выбирают прогнозируемый класс, который максимизирует апостериорное распределение вероятностей; разница только в подходе.

Логистическая регрессия

Одним из популярных алгоритмов классификации является логистическая регрессия, и его часто спрашивают в сочетании с линейной регрессией во время интервью как способ оценить базовые знания об алгоритмах классификации. В логистической регрессии мы берем линейный результат и преобразуем его в вероятность от 0 до 1, используя сигмовидную функцию:

$$S(x) = \frac{1}{1 + e^{-x}}$$

В матричной форме решение выглядит следующим образом, где 1 — это целевой класс, если результат не менее 0,5:

$$P(\hat{Y} = 1|x) = S(w^T x)$$

Функция потерь для логистической регрессии представляет собой логарифмическую потерю:

$$L(w) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{S(w^T x)}\right) + (1 - y_i)\left(\frac{1}{1 - S(w^T x)}\right)$$

Обратите внимание, что апостериорная модель моделируется напрямую, и, следовательно, логистическая регрессия является дискриминационной моделью.

Линейный дискриминантный анализ

Линейный дискриминантный анализ (LDA) не является часто задаваемым вопросом во время интервью, но служит интересной темой для изучения, поскольку это генеративная модель, а не дискриминативная модель (которой была логистическая регрессия). Предполагается, что для некоторого класса k распределение любых данных из этого класса следует многомерному гауссовскому закону:

$$X|Y = k \sim N(\mu_k, \Sigma_k)$$

Напомним из правила Байеса, что максимизация совместной вероятности по меткам эквивалентна максимизации апостериорной вероятности, поэтому LDA стремится максимизировать:

$$\hat{y} = \text{argmax}_k p(Y = k|x)$$

В частности, у нас есть:

$$P(Y = k|x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^{K} f_i(x)\pi_i}$$

где f(x) для каждого k — функция плотности классов. LDA предполагает, что плотности являются многомерными гауссовыми, и дополнительно предполагает, что ковариационная матрица является общей для всех классов. Результирующая граница решения является линейной (отсюда и название), поскольку существует также квадратичный дискриминантный анализ, где граница является квадратичной.

Деревья решений

Во время интервью обычно задают вопросы о деревьях решений и случайных лесах, поскольку они представляют собой гибкие и часто хорошо работающие модели на практике. В частности, это помогает иметь общее представление о том, как оба обучаются и используются, а также о том, как происходит разделение функций (энтропия и прирост информации).

Тренировочные деревья решений

Дерево решений — это модель, которая может быть представлена в виде дерева, при этом при каждом разбиении происходит разделение на основе признаков, что приводит к различным листовым узлам, в результате чего возникает результат (классификация или регрессия). В этом обсуждении мы сосредоточимся на настройке классификации. Они обучаются жадным и рекурсивным образом, начиная с корня, где цель состоит в том, чтобы выбрать разбиения, которые повышают наибольшую уверенность в том, к какому классу принадлежит конкретная точка данных.

Энтропия

Энтропия случайной величины Y количественно определяет неопределенность ее значений и определяется следующим образом для дискретной переменной Y, которая принимает k состояний:

$$H(Y) = -\sum_{i=1}^{k} P(Y = k) \log P(Y = k)$$

Для простой случайной величины Бернулли эта величина максимальна при p = 0,5 и минимальна при p = 0 или p = 1, что интуитивно согласуется с определением, поскольку при p = 0 или 1 неопределенность результата отсутствует. Как правило, если случайная величина имеет высокую энтропию, то ее распределение ближе к равномерному, чем к асимметричному.

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

Рассмотрим произвольное разделение. У нас есть H(Y) с начальных обучающих меток, и, скажем, у нас есть некоторая функция X, которую мы хотим разделить. Мы можем охарактеризовать снижение неопределенности информационным приростом, который определяется как:

$$IG(Y, X) = H(Y) - H(Y|X)$$

Чем больше это количество, тем выше снижение неопределенности в Y за счет разделения на X. Следовательно, общий процесс заключается в оценке всех рассматриваемых признаков и выборе признака, который максимизирует этот прирост информации. Затем рекурсивно продолжите процесс для двух полученных ветвей.

Случайные леса

Как правило, отдельное дерево решений может быть склонно к переоснащению, поэтому на практике обычно случайные леса дают лучшие прогнозы вне выборки. Случайный лес — это метод ансамбля, который использует множество деревьев решений и усредняет решение из них. Это уменьшает переобучение и корреляцию между деревьями двумя методами: 1) создание пакетов (агрегация начальной загрузки), посредством чего некоторые точки данных m < n (где n — общее количество) произвольно выбираются с заменой и используются в качестве обучающего набора, 2) случайное подмножество функций рассматривается при каждом разделении (чтобы предотвратить постоянное разделение по какой-либо конкретной функции).

Кластеризация

Кластеризация — популярная тема на собеседованиях, поскольку существует множество реальных приложений. Это часто делается для визуализации данных и может использоваться для выявления выбросов, которые полезны в таких случаях, как обнаружение мошенничества. Это также помогает иметь общее представление о том, как параметры изучаются в этом контексте, по сравнению с подходом MLE/MAP из предыдущего.

Обзор кластеризации

Цель кластеризации состоит в том, чтобы разделить набор данных на различные кластеры, рассматривая только входные функции. Это пример обучения без учителя. В идеале кластеризация имеет два свойства: 1) точки внутри данного кластера похожи друг на друга (высокое внутрикластерное сходство).

2) точки в разных кластерах не похожи друг на друга (низкое межкластерное сходство).

Кластеризация K-средних

Кластеризация K-средних разбивает данные на k кластеров и начинается с произвольного выбора

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

центроидов каждого из k кластеров. Итеративно он обновляет разделы, назначая точки ближайшему кластеру, обновляя центроиды и повторяя до конвергенции.

Математически K-means решает следующую проблему, минимизируя следующую функцию потерь (данные точки и значения центроида):

$$L = \sum_{j=1}^{k} \sum_{x \in S_j} || \, x_i - \mu_j \, ||^2$$

Итеративный процесс продолжается до тех пор, пока обновления назначения кластера не улучшат целевую функцию.

Модель гауссовой смеси

Модель гауссовой смеси (GMM) — это модель, в которой для любой заданной точки данных x мы предполагаем, что она исходит из одного из k кластеров, каждый из которых имеет определенное распределение Гаусса.

То есть среди K классов имеем:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

где коэффициенты π представляют собой коэффициенты смешивания в кластерах и нормализованы, поэтому их сумма равна 1. Пусть θ обозначает неизвестное среднее значение и параметры дисперсии для каждого из K классов, а K — коэффициенты смешивания. Тогда вероятность определяется как:

$$p(\theta \mid X) = \prod_{i=1}^{n} p(x) = \prod_{i=i}^{n} \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

и, следовательно, логарифмическая вероятность:

$$\log p(\theta \mid X) = \sum_{i=i}^{n} \log \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

Параметры могут быть рассчитаны итеративно с использованием максимизации ожидания (EM), которая обсуждается ниже.

Максимизация ожиданий

Максимизация ожидания (EM) — это метод оценки параметров для скрытых переменных, таких как два приведенных выше примера K-средних и GMM, при котором некоторые переменные можно наблюдать напрямую, тогда как другие являются скрытыми и не могут наблюдаться напрямую. В частности, для кластеризации назначение кластера является скрытой переменной, поскольку оно не наблюдается напрямую. Общие шаги заключаются в следующем: Z используется как скрытые переменные, X как наблюдаемые переменные и неизвестные параметры θ. Предположим,

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

что текущие параметры задаются как: θ'. Первым шагом является оценка:

$$p(Z|X,\theta')$$

используя текущие оценки параметров. Второй шаг заключается в оценке наиболее вероятного θ*, который максимизирует логарифмическую вероятность данных, которая определяется как:

$$\sum_{Z} p(Z \mid X,\theta') \log p(X,Z \mid \theta)$$

И так далее до сходимости.

30 задач на собеседовании по машинному обучению

18 вопросов средней сложности для собеседования по машинному обучению

1. **(Робинхуд)** Что такое отток пользователей и как можно построить модель, чтобы предсказать, будет ли отток пользователей? Какие функции вы бы включили в модель и как вы оцениваете их важность?

2. **(Подтвердить)** Предположим, у нас есть классификатор, который дает оценку от 0 до 1 для вероятности того, что конкретная заявка на получение кредита является мошеннической. В этом сценарии: а) что такое ложноположительные результаты, б) что такое ложноотрицательные результаты и в) каковы компромиссы между ними в долларовом выражении и как следует соответственно взвешивать модель?

3. **(Uber)** Допустим, вам нужно создать бинарный классификатор для обнаружения мошенничества. Какие метрики вы бы рассмотрели, как они определяются и как их интерпретируют?

4. **(Google)** Допустим, вам дан очень большой набор слов. Как бы вы определили синонимы?

5. **(Airbnb)** Допустим, вы проводите простую логистическую регрессию для решения проблемы, но результаты вас не удовлетворяют. Какими способами вы могли бы улучшить свою модель или какие другие модели вы могли бы изучить?

6. **(Amazon)** Опишите как генеративную, так и дискриминативную модели и приведите пример каждой из них?

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google. Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ
ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ
ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL
14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ
ОБО МНЕ

7. **(Подтвердить)** Предположим, что у нас есть некоторая модель кредитоспособности, в которой точно откалибрована (с точностью до некоторой погрешности) оценка того, насколько кредитоспособен любой отдельный человек. Например, если оценка модели составляет 92%, мы можем предположить, что фактическая оценка находится между 91 и 93. Если мы возьмем 92 в качестве порога оценки и посчитаем всех, кто выше этого балла, достойными кредита, переоцениваем ли мы или недооцениваем кредитный рейтинг фактического населения?

8. **(Microsoft)** Каков компромисс между смещением и дисперсией? Как это выразить с помощью уравнения?

9. **(Uber)** Определение процесса перекрестной проверки. Какова мотивация его использования?

10. **(Airbnb)** Допустим, вы моделируете годовой доход от новых объявлений. Какие функции вы бы использовали? Какие шаги обработки данных необходимо предпринять и какая модель будет работать?

11. **(Stitch Fix)** Как бы вы построили модель для расчета склонности клиента к покупке определенного товара? Каковы некоторые плюсы и минусы вашего подхода?

12. **(Uber)** Что такое регуляризация L1 и L2? Каковы различия между ними?

13. **(Amazon)** Определите, что означает выпуклость функции. Приведите пример алгоритма машинного обучения, который не является выпуклым, и объясните, почему это так?

14. **(Подтвердить)** Предположим, у нас есть классификатор, который дает оценку от 0 до 1 для вероятности того, что конкретная заявка на получение кредита стоит за мошенничеством. Предположим, что для оценки каждого приложения мы извлекаем квадратный корень из этой оценки. Как изменится кривая ROC? Если она не изменится, какие функции изменят кривую?

15. **(Amazon)** Опишите градиентный спуск и мотивы стохастического градиентного спуска?

16. (**Майкрософт)** Объясните, что такое прирост информации и энтропия в дереве решений?

17. **(Airbnb)** Предположим, вам поручили создать

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

модель, которая может рекомендовать похожие объявления пользователям Airbnb, когда они просматривают любое данное объявление. Какую модель вы бы использовали, какие данные необходимы для этой модели и как бы вы оценили модель?

12 сложных вопросов на собеседовании по машинному обучению

18. **(Microsoft)** Опишите идею бустинга. Приведите пример одного метода и опишите одно его преимущество и недостаток?

19. **(Google)** Допустим, мы запускаем вероятностную линейную регрессию, которая хорошо моделирует лежащую в основе взаимосвязь между некоторыми y и x. Теперь предположим, что ко всем входным данным добавлен некоторый шум ε, который не зависит от обучающих данных. Какова новая целевая функция? Как вы это вычисляете?

20. **(Netflix)** Какая функция потерь используется в кластеризации k-средних для k кластеров и n точек выборки? Вычислите формулу обновления, используя 1) пакетный градиентный спуск, 2) стохастический градиентный спуск для среднего значения кластера для кластера k с использованием скорости обучения ε.

21. **(Tesla)** Вы работаете с несколькими датчиками, которые предназначены для прогнозирования определенного показателя энергопотребления автомобиля. Используя выходные данные датчиков, вы строите модель линейной регрессии, чтобы сделать прогноз. Датчиков много, и некоторые из них склонны к полному отказу. Какие функции затрат вы могли бы рассмотреть и какие из них вы решили минимизировать в этом сценарии?

22. **(Полоса)** Допустим, мы используем смешанную модель Гаусса (GMM) для обнаружения аномалий в мошеннических транзакциях, чтобы классифицировать входящие транзакции по K классам. Формулно опишите настройку модели и как оценить апостериорные вероятности и логарифмическую вероятность. Как мы можем определить, следует ли считать новую транзакцию мошеннической?

23. **(Netflix)** Что такое максимизация ожиданий и когда она полезна? Алгоритмически опишите установку с помощью формул.

24. **(Opendoor)** Опишите установку и допущения

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

использования линейного дискриминантного анализа (LDA). Покажите математически, что границы решений линейны.

25. **(Microsoft)** Сформулируйте предысторию SVM и покажите проблему оптимизации, которую она призвана решить.

26. **(Netflix)** Опишите энтропию в контексте машинного обучения и математически покажите, как ее максимизировать, предполагая N состояний.

27. **(Airbnb)** Предположим, вы используете линейную регрессию и моделируете ошибки как нормально распределенные. Покажите, что в этой установке максимизация правдоподобия данных эквивалентна минимизации суммы квадратов остатков.

28. **(Полоса)** Опишите формулировку модели, лежащую в основе логистической регрессии. Как максимизировать логарифмическую вероятность данной модели (используя случай двух классов)?

29. **(Netflix)** Скажем, X — одномерная гауссовская случайная величина. Чему равна энтропия X?

30. **(Uber)** Опишите идею PCA и проанализируйте ее формулировку и вывод в матричной форме. Пройдите процедурное описание и решите задачу максимизации с ограничениями.

Если вы жаждете новых задач, ознакомьтесь с 40 задачами на собеседовании по проблемам и статистике и нашим бесплатным 9-дневным экспресс-курсом на собеседование по науке о данных .

8 решений для интервью с машинным обучением

Проблема №3 Решение:

Некоторыми основными важными из них являются точность, полнота и AUC кривой ROC. Определим TP как истинно положительный, FP как ложноположительный, TN как истинно отрицательный и FN как ложноотрицательный.

Точность определяется TP / (TP + FP). Он отвечает на вопрос «какой процент мошеннических прогнозов оказался верным?» и важно максимизировать, поскольку вы хотите, чтобы ваш классификатор был максимально точным, когда он идентифицировал транзакцию как мошенническую.

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ

Отзыв определяется TP / (TP + FN). Он отвечает на вопрос «какой процент случаев мошенничества был выявлен?» и важно максимизировать, так как вы хотите, чтобы ваш классификатор поймал как можно больше случаев мошенничества.

AUC ROC-кривой определяется площадью под ROC-кривой, которая строится путем построения графика частоты истинных положительных результатов TP/(TP+FN) по сравнению с частотой ложноположительных результатов FP/(FP+TN) для различных пороговых значений, которые определяет ярлык мошеннических или не мошеннических. Площадь под этой кривой, или AUC, является мерой отделимости модели, и чем она ближе к 1, тем выше мера. Он отвечает на вопрос «способен ли мой классификатор эффективно различать мошенничество и не мошенничество», и его важно максимизировать, поскольку вы хотите, чтобы ваш классификатор соответствующим образом разделял два класса.

Проблема №8 Решение:

Уравнение, в котором этот компромисс выражается, дается следующим образом: Общая ошибка модели = Смещение + Дисперсия + Неустранимая ошибка. Гибкие модели имеют низкое смещение и высокую дисперсию, тогда как более жесткие модели имеют высокое смещение и низкую дисперсию.

Термин смещения возникает из-за ошибки, когда модель не соответствует данным. Наличие высокого смещения означает, что модель машинного обучения слишком проста и не отражает взаимосвязь между функциями и целью. Примером может служить линейная регрессия, когда базовая связь нелинейна.

Термин дисперсии возникает из-за ошибки, когда модель переобучает данные. Наличие высокой дисперсии означает, что модель восприимчива к изменениям обучающих данных, что означает, что она улавливает слишком много шума. Примером может служить очень сложная нейронная сеть, в которой истинная базовая связь между функциями и целью является линейной.

Неустранимый термин возникает из-за ошибки, которая не может быть устранена непосредственно моделью, например, из-за шума в измерениях данных.

Задача №12 Решение:

Регуляризация L1 и L2 — это оба метода регуляризации, которые пытаются предотвратить переобучение в машинном обучении. Для обычной модели регрессии предположим, что функция потерь задается L. L1 добавляет абсолютное значение коэффициентов в качестве штрафного члена, тогда как L2 добавляет квадрат величины коэффициентов в качестве штрафного члена.

Функция потерь для них:

$$Loss(L_1) = L + \lambda|w_i|$$

$$Loss(L_2) = L + \lambda|w_i^2|$$

Где функция потерь L представляет собой сумму квадратов ошибок, определяемую следующим образом, где f(x) представляет собой интересующую модель, например, линейную регрессию с предикторами p:

$$L = \sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}(x_{ij}w_j))^2 \text{ for linear regression}$$

Если мы запустим градиентный спуск для весов w, мы обнаружим, что регуляризация L1 заставит любой вес приблизиться к 0, независимо от его величины, тогда как для регуляризации L2 скорость, с которой вес приближается к 0, становится медленнее, чем скорость увеличивается. к 0. Из-за этого L1 с большей вероятностью «обнуляет» определенные веса и, следовательно, полностью удаляет определенные функции из модели, что приводит к более разреженным моделям.

Задача №15 Решение:

Градиентный спуск — это алгоритм, который делает небольшие шаги в направлении наискорейшего спуска для конкретной целевой функции. Скажем, у нас есть функция f(), и мы находимся в какой-то точке x в момент времени t. Затем градиентный спуск будет обновлять x следующим образом до сходимости:

$$x^{t+1} = x^t - \alpha_t \nabla f(x^t)$$

то есть мы вычисляем отрицательное значение градиента f и масштабируем его на некоторую константу и двигаемся в этом направлении на каждой итерации.

Поскольку многие функции потерь можно разложить на сумму отдельных функций, то общий шаг градиента можно разбить на добавление отдельных градиентов. Однако для очень больших наборов данных это может потребовать значительных вычислительных ресурсов, и алгоритм может застрять в локальных минимумах или седловых точках.

Следовательно, мы можем использовать стохастический градиентный спуск (SGD), в котором мы получаем несмещенную оценку истинного градиента, не проходя через все точки данных, путем равномерного случайного выбора точки и выполнения там обновления градиента.

Оценка несмещена, так как имеем:

$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x)$$

и поскольку предполагается, что данные являются iid, то для SGD g(x) в ожидании:

$$E[g(x)] = \nabla f(x)$$

19. Вспомним целевую функцию для линейной регрессии, где x — набор входных векторов, а w — веса:

$$L(w) = E[(w^T x - y)^2]$$

Предположим, что добавленный шум является гауссовым следующим образом:

$$\epsilon \sim N(0, \lambda I)$$

Тогда новая целевая функция имеет вид:

$$L(w) = E[(w^T (x + \epsilon) - y)^2]$$

Чтобы вычислить его, мы упрощаем:

$$L'(w) = E[(w^T x - y + w^T \epsilon)^2]$$

$$L'(w) = E[(w^T x - y)^2 + 2(w^T x - y) w^T \epsilon + w^T \epsilon \epsilon^T w]$$

$$L'(w) = E[(w^T x - y)^2] + E[2(w^T x - y) w^T \epsilon] + E[w^T \epsilon \epsilon^T w]$$

Мы знаем, что ожидание для ε равно 0, поэтому средний член становится равным 0, и у нас остается:

$$L'(w) = L(w) + 0 + w^T E[\epsilon \epsilon^T] w$$

Последний термин можно упростить следующим образом:

$$L'(w) = L(w) + w^T \lambda I w$$

И поэтому целевая функция упрощается до L2-регуляризации:

$$L'(w) = L(w) + \lambda ||w||^2$$

Задача №21 Решение:

Здесь есть две функции потенциальных затрат: одна использует норму L1, а другая использует норму L2. Ниже приведены две основные функции стоимости, использующие нормы L1 и L2 соответственно:

$$J(w) = ||Xw - y||$$

$$J(w) = |Xw - y|^2$$

В этом случае было бы разумнее использовать норму L1, поскольку норма L1 сильнее наказывает выбросы и, таким образом, придает меньшее значение полным отказам, чем норма L2.

Кроме того, было бы разумно включить член регуляризации для учета шума. Если предположить, что шум добавляется к каждому датчику равномерно следующим образом:

$$\epsilon \sim N(0, \lambda I)$$

тогда, используя традиционную регуляризацию L2, мы получим функцию стоимости:

$$J(w) = ||Xw - y|| + \lambda ||w||^2$$

Однако, учитывая тот факт, что существует множество датчиков (и широкий диапазон их полезности), мы могли бы вместо этого предположить, что шум добавляется:

$$\epsilon \sim N(0, \lambda D)$$

где каждый диагональный член в матрице D представляет член ошибки, используемый для каждого датчика (и, следовательно, штрафует одни датчики больше, чем другие). Тогда наша окончательная функция стоимости определяется как:

$$J(w) = ||Xw - y|| + \lambda w^T D w$$

Задача № 25 Решение :

Цель SVM — сформировать гиперплоскость, которая линейно разделяет заданные обучающие данные. В частности, он направлен на максимизацию запаса, который представляет собой минимальное расстояние от границы решения до любой точки обучения. Ближайшие к гиперплоскости точки называются опорными векторами.

Математически гиперплоскость задается следующим образом для некоторой константы c:

$$H = \{h : w^T h = c\}$$

Теперь рассмотрим некоторую произвольную точку x_i, которая не лежит на гиперплоскости. Расстояние от x_i до H — это длина проекции от x_i-h до вектора, перпендикулярного H:

$$d = \frac{|w^T(x_i - h)|}{||w||_2} = \frac{|w^T x_i - c|}{||w||_2}$$

Чтобы получить фактические признаки классификации (положительные или отрицательные), мы можем умножить это расстояние на знак для каждого y_i:

$$y_i * \frac{(w^T x_i - c)}{||w||_2}$$

Если предположить, что запас равен m, то задача оптимизации состоит в следующем:

$$\max m \text{ s.t. } y_i * \frac{(w^T x_i - c)}{||w||_2} \geq m$$

Для обеспечения уникальности мы можем установить ограничение на m:

$$m = \frac{1}{||w||_2}$$

И поэтому окончательная задача оптимизации:

$$\max \frac{1}{||w||_2} \text{ s.t. } y_i * (w^T x_i - c) \geq 1$$

Задача № 29 Решение :

У нас есть:

$$X \sim N(\mu, \sigma^2)$$

а энтропия для непрерывной случайной величины определяется выражением:

$$H(x) = -\int_{-\}^{\infty}p(x)\log p(x) dx$$

Для гауссиана имеем:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)}{2\sigma^2}}$$

Подключение к вышеперечисленным дает:

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log \sigma\sqrt{2\pi} dx - \int_{-\infty}^{\infty} p(x)(-\frac{(x-\mu)}{2\sigma^2})dx$$

Первый член равен

$$-\log \sigma\sqrt{2\pi} \int_{-\infty}^{\infty} p(x)dx = -\log \sigma\sqrt{2\pi}$$

поскольку интеграл равен 1 (по определению плотности вероятности). Второй термин определяется:

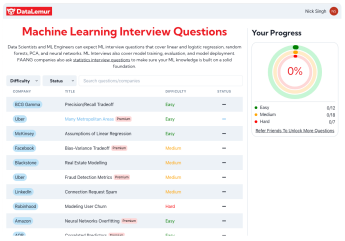
$$\frac{1}{2\sigma^2} \int_{-\infty}^{\infty} p(x)(x-\mu)^2 dx = \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}$$

поскольку внутренний член является выражением для дисперсии. Следовательно, энтропия равна:

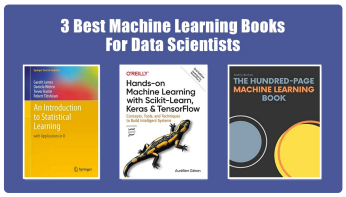
$$H(x) = \frac{1}{2} - \log \sigma\sqrt{2\pi}$$

Где взять больше вопросов для интервью по науке о данных

Хотите больше подобного? Купите Ace the Data Science Interview на Amazon Prime и попрактикуйтесь в вопросах для интервью по машинному обучению на DataLemur !



Вам также понравится этот список лучших книг для специалистов по данным , в котором представлены 3 наши любимые книги по машинному обучению.



Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

Ник Сингх

Автор бестселлера Amazon, автор интервью Ace the Data Science и создатель курса Ace the Data Job Hunt.

Основатель DataLemur и ранее инженер-программист в Facebook и Google.
Присоединяйтесь к моему бесплатному 9-дневному ускоренному курсу Data Interview!

БЛОГ

ИНТЕРВЬЮ С ЭЙСОМ О НАУКЕ О ДАННЫХ

ПОДГОТОВКА К СОБЕСЕДОВАНИЮ DATALEMUR ПО SQL

14 КНИГ, КОТОРЫЕ ИЗМЕНИЛИ МОЮ ЖИЗНЬ

ОБО МНЕ