

Оглавление

DECIDING WHAT TO TRY NEXT.....	1
EVALUATING A HYPOTHESIS.....	3
MODEL SELECTION AND TRAIN VALIDATION TEST SETS.....	5

DECIDING WHAT TO TRY NEXT

0:00

К настоящему времени вы видели множество различных алгоритмов обучения.

0:03

И если вы следовали этим видео, вы должны считать себя экспертом по многим современным технологиям машинного обучения. Но даже среди людей, которые знают определенный алгоритм обучения. Часто существует огромная разница между тем, кто действительно знает, как эффективно и эффективно применять этот алгоритм, по сравнению с тем, кто менее знаком с некоторыми материалами, которые я собираюсь учить, и которые на самом деле не понимают, как применять эти алгоритмы, и могут в конечном итоге тратить много времени, пытаясь разобраться, что на самом деле не имеет смысла.

0:34

Я хотел бы сделать это, чтобы убедиться, что, если вы разрабатываете системы машинного обучения, вы знаете, как выбрать один из самых перспективных способов провести свое время. И в этом и следующих нескольких видео я собираюсь дать ряд практических советов, рекомендаций, как это сделать. И конкретно, на что мы будем сосредоточены, проблема в том, что вы разрабатываете систему машинного обучения или пытаетесь улучшить производительность системы машинного обучения, как вы решаете, какие прокси-серверы пытаются попробовать

1:07

следующий?

1:09

Чтобы объяснить это, давайте продолжим использовать наш пример обучения для прогнозирования цен на жилье. И предположим, что вы реализуете и регулируете линейную регрессию. Таким образом, минимизируя эту функцию стоимости J . Теперь предположим, что после того, как вы возьмете свои параметры обучения, если вы проверите свою гипотезу о новом наборе домов, предположите, что вы обнаружите, что это приводит к огромным ошибкам в этом прогнозе цен на жилье.

1:33

Вопрос в том, что вы должны попробовать, чтобы улучшить алгоритм обучения?

1:39

Есть много вещей, о которых можно подумать, что может улучшить производительность алгоритма обучения.

1:44

Одна вещь, которую они могли бы попробовать, - это получить больше примеров обучения. И конкретно, вы можете себе представить, может быть, вы знаете, настраивая телефонные опросы, идя от двери до двери, чтобы попытаться получить больше данных о том, сколько разных домов продают.

1:57

И печально то, что я видел много людей, которые тратят много времени на сбор дополнительных примеров обучения, думая о, если у нас в два раза больше или в десять раз больше данных о тренировках, это, безусловно, поможет, верно? Но иногда получение большего количества учебных данных на самом деле не помогает, и в следующих нескольких видео мы увидим, почему, и мы увидим, как вы можете не тратить много времени на сбор дополнительных данных обучения в настройках, где он просто не поможет.

2:22

Другие вещи, которые вы можете попробовать, - это, возможно, попробовать меньший набор функций. Поэтому, если у вас есть набор функций, таких как x_1 , x_2 , x_3 и т. Д., Возможно, большое количество функций. Возможно, вы хотите провести время, тщательно подбирая некоторые небольшие подмножества, чтобы предотвратить переобучение.

2:38

Или, может быть, вам нужно получить дополнительные функции. Возможно, текущий набор функций недостаточно информативен, и вы хотите собрать больше данных в смысле получения дополнительных функций.

2:48

И еще раз это проект, который может расширить масштабы огромных проектов. Можете ли вы представить себе опросы по телефону, чтобы узнать больше о домах или провести дополнительные обследования земель, чтобы узнать больше о земельных участках и т. Д., Так что это огромный проект. И еще раз было бы интересно узнать заранее, если это поможет, прежде чем мы потратим много времени на то, чтобы сделать что-то подобное. Мы также можем попробовать добавить многочленные функции, такие как x_2^2 и функции продукта $x_1 x_2$. Мы все еще можем потратить немало времени на размышления об этом, и мы также можем попробовать другие вещи, такие как уменьшение λ , параметр регуляризации или увеличение λ .

3:23

Учитывая меню таких параметров, некоторые из них могут легко масштабировать до шести месяцев или более проектов.

3:31

К сожалению, самый распространенный метод, который люди используют, чтобы выбрать один из них, - это почувствовать чувство кишки. В чем то, что многие люди будут делать, это случайный выбор одного из этих вариантов и, возможно, сказать: «О, отпустите и получите больше данных о тренировках». И легко потратите шесть месяцев на сбор дополнительных данных о тренировках, или, может быть, кто-то еще лучше скажет: «Ну, пойдём, собираем намного больше возможностей в этих домах в нашем наборе данных». И у меня много раз, к сожалению, люди тратят, знаете ли, буквально 6 месяцев, выполняя один из этих способов, что они могут научиться случайным образом только через шесть месяцев, что это действительно не было перспективным направлением.

4:07

К счастью, есть довольно простой метод, который может позволить вам очень быстро исключить половину вещей в этом списке как потенциально перспективные вещи для продолжения. И есть очень простая методика, которая, если вы запустите, может легко исключить многие из этих опций и потенциально сэкономить вам много времени на то, что просто не сработает.

4:29

В следующих двух видео после этого я расскажу о том, как оценивать алгоритмы обучения.

4:36

И в следующих нескольких видео после этого я расскажу об этих методах, которые называются диагностикой машинного обучения.

4:46

И что такое диагностика, это тест, который вы можете запустить, чтобы понять, что есть или не работает с алгоритмом, и который часто дает вам представление о том, какие перспективные вещи пытаются улучшить алгоритм обучения

5:03

представление. В этой видеопоследовательности мы поговорим об определенной диагностике. Но я должен заранее указать, что диагностика может потребовать времени для реализации, и иногда, как вы знаете, требуется довольно много времени для реализации и понимания, но это может очень хорошо использовать ваше время, когда вы разрабатываете алгоритмы обучения, потому что они часто могут сэкономить вам много времени, проводя протест, который вы могли бы узнать намного раньше, просто не будет плодотворным.

5:32

Поэтому в следующих нескольких видеороликах я сначала расскажу о том, как оценивать ваши алгоритмы обучения, и после этого я расскажу о некоторых из этих диагностических программ, которые, надеюсь, позволят вам гораздо более эффективно выбирать больше полезных вещей, чтобы попробовать если ваша цель улучшить систему машинного обучения.

[EVALUATING A HYPOTHESIS](#)

0:00

В этом видео я хотел бы рассказать о том, как оценить гипотезу, которая была изучена вашим алгоритмом. В более поздних видеороликах мы остановимся на этом, чтобы поговорить о том, как предотвратить проблемы с переобучением и недоукомплектованием. Когда мы подходим к параметрам нашего алгоритма обучения, мы думаем о выборе параметров для минимизации ошибки обучения. Можно подумать, что получение действительно низкой стоимости ошибки обучения может быть хорошей вещью, но мы уже видели это только потому, что гипотеза имеет низкую ошибку обучения, это не значит, что она обязательно является хорошей гипотезой. И мы уже видели пример того, как гипотеза может переработать. И поэтому не могут обобщать новые примеры не в учебном наборе. Итак, как вы скажете, может ли гипотеза перефитировать. В этом простом примере мы можем построить гипотезу h of x и просто посмотреть, что происходит. Но в целом для проблем с большим количеством функций, чем только одна особенность, для проблем с большим количеством таких функций становится трудно или невозможно построить сюжет, как выглядит гипотеза, и поэтому нам нужен другой способ оценить нашу гипотезу. Стандартный способ оценки усвоенной гипотезы заключается в следующем. Предположим, что у нас есть набор данных, подобный этому. Здесь я только что показал 10 учебных примеров, но, конечно, обычно у нас могут быть десятки или сотни или, может быть, тысячи примеров обучения. Чтобы убедиться, что мы можем оценить нашу гипотезу, что мы собираемся сделать, это разделить данные, которые мы имеем на две части. Первая часть будет нашим обычным набором тренировок

1:42

и вторая часть будет нашим тестовым набором, и довольно типичный раскол этого, все данные, которые мы имеем в наборе для обучения и наборе тестов, могут быть примерно равны 70%, 30%. Стоит больше сегодня оценивать учебный комплект и относительно меньше, чем набор тестов. И теперь, если у нас есть некоторый набор данных, мы используем синус, скажем, 70% данных, чтобы быть нашим учебным комплексом, где здесь « m », как обычно, число наших примеров обучения, а остальные наши данные могут быть назначены чтобы стать нашим тестовым набором. И здесь я собираюсь использовать тест индексов m для обозначения числа тестовых примеров. И, как правило, этот тест под индексом будет означать примеры, которые исходят из тестового набора, так что тест подстроки x_1 , тест подстроки y_1 - это мой первый тестовый пример, который, как мне кажется, в этом примере может быть здесь. Наконец, последняя деталь, тогда как здесь я нарисовал это, как будто первые 70% идут на тренировочный набор и последние 30% на тестовый набор. Если есть какие-то обычные данные. Это должно быть лучше отправить случайные 70% ваших данных в набор обучения и случайные 30% ваших данных в тестовый набор. Поэтому, если ваши данные уже были отсортированы случайным образом, вы можете просто взять первые 70% и последние 30%, если бы ваши данные не были упорядочены произвольно, было бы лучше случайным образом перетасовать или случайным образом переупорядочить примеры в вашем учебном наборе. Прежде чем вы узнаете отправку первых 70% в тренировочном наборе и последние 30% набора тестов. Вот тогда довольно типичная процедура для того, как вы будете тренировать и тестировать алгоритм обучения и регрессию обучения. Во-первых, вы изучите параметры θ из набора тренировок, чтобы свести к минимуму обычную цель ошибки обучения $J(\theta)$, где $J(\theta)$ здесь был определен, используя 70% всех данных, которые у вас есть. Имеются только данные обучения. И тогда вы вычислили ошибку теста. И я собираюсь обозначить тестовую ошибку как J индексный тест. Итак, что вы делаете, это

взять свой параметр θ , который вы узнали из учебного набора, и подключить его сюда и вычислить ошибку теста. Что я буду писать следующим образом. Таким образом, это в основном средняя квадратичная ошибка, измеренная на вашем тестовом наборе. Это в значительной степени то, что вы ожидаете. Поэтому, если мы запускаем каждый тестовый пример с помощью вашей гипотезы с параметром θ и просто измеряем квадратичную ошибку, которую ваша гипотеза имеет на вашем тестовом примере m , тестовые примеры. И, конечно же, это определение ошибки тестового набора, если мы используем линейную регрессию и используем квадратичную метрику ошибки. Как насчет того, если мы будем заниматься проблемой классификации и скажем, используя логистическую регрессию. В этом случае процедура обучения и тестирования говорит, что логистическая регрессия довольно схожа, сначала мы сделаем параметры из данных тренинга, что первые 70% данных. И он вычислит ошибку теста следующим образом. Это та же целевая функция, что и мы всегда используем, но мы просто логистическая регрессия, за исключением того, что теперь определяется с помощью нашего теста подстроки m , тестовых примеров. Хотя это определение теста тестового набора j подстроичный тест вполне разумно. Иногда есть альтернативная метка тестов, которая может быть легче интерпретировать, и это ошибка ошибочной классификации. Он также называется нулевой ошибкой ошибочной классификации, с нулевым значением, означающим, что вы либо получаете пример вправо, либо получаете неправильный пример. Вот что я имею в виду. Позвольте мне определить ошибку предсказания.

Это $h(x)$. И учитывая, что метка y равна единице, если моя гипотеза выводит значение, большее, чем пять, а y равно нулю или если моя гипотеза выводит значение меньше 0,5, а y равно единице, так что обе эти в случаях, если ваша гипотеза неверно маркирует пример, предполагающий ваш порог в 0,5. Таким образом, либо считалось, что это скорее всего 1, но на самом деле это 0, или ваша гипотеза хранится с большей вероятностью 0, но метка была фактически 1. Иначе мы определяем эту функцию ошибки равной нулю. Если ваша гипотеза в основном классифицировала пример y правильно. Затем мы могли бы определить ошибку теста, используя метрику ошибки ошибочной классификации, чтобы быть одним из m тестов суммы из i , равным одному-м индексу теста ошибки $h(x(i))$ тестовой запятой $y(i)$. И это только мой способ написать, что это именно та часть примеров в моем тестовом наборе, что моя гипотеза неверно маркирована. И так это определение ошибки тестового набора, используя ошибку ошибочной классификации метрики классификации 0 1. Итак, это стандартная методика оценки того, насколько хороша научная гипотеза. В следующем видео мы адаптируем эти идеи, чтобы помочь нам делать такие вещи, как выбирать, какие функции, такие как полином степени, который можно использовать с алгоритмом обучения, или выбрать параметр регуляризации для алгоритма обучения.

MODEL SELECTION AND TRAIN VALIDATION TEST SETS

0:00

Предположим, вам осталось решить, какую степень полинома соответствует набору данных. Итак, какие функции включать, что дает вам алгоритм обучения. Или предположим, что вы хотите выбрать параметр регуляризации больше для алгоритма обучения. Как ты это делаешь? Этот

процесс выбора модели учетной записи. Браузеры, и в нашем обсуждении, как это сделать, мы поговорим не только о том, как разделить ваши данные на поездках и наборах тестов, но как переключить данные в то, что мы обнаруживаем, называется набором поездов, валидации и испытаний, Мы увидим в этом видео только то, что это такое, и как использовать их для выбора модели. Мы уже много раз видели проблему переобучения, в которой только потому, что алгоритм обучения хорошо подходит для тренировочного набора, это не значит, что это хорошая гипотеза. В более общем плане, именно поэтому ошибка обучающего набора не является хорошим предиктором того, насколько гипотеза будет действовать на новом примере. Конкретно, если вы соответствуете некоторому набору параметров. $\theta_0, \theta_1, \theta_2$ и т. Д., К вашему тренировочному набору. Тогда тот факт, что ваша гипотеза преуспевает в учебном наборе. Ну, это не означает многого с точки зрения прогнозирования того, насколько хорошо ваша гипотеза будет обобщаться на новые примеры, не замеченные в учебном наборе. И более общий принцип заключается в том, что как только ваш параметр подходит для некоторого набора данных. Может быть, тренировочный набор, может быть, что-то еще. Тогда ошибка вашей гипотезы, измеренная в том же наборе данных, например, ошибка обучения, вряд ли будет хорошей оценкой вашей фактической ошибки обобщения. Именно так гипотеза будет обобщаться на новые примеры. Теперь рассмотрим проблему выбора модели. Предположим, вы пытаетесь выбрать, какой степенной полином подходит для данных. Итак, вы должны выбрать линейную функцию, квадратичную функцию, кубическую функцию? Весь путь до полинома 10-го порядка.

1:51

Так что, как будто в этом алгоритме есть один дополнительный параметр, который я буду обозначать d , а именно, какую степень полинома. Вы хотите выбрать. Так что, как и в θ -параметрах, это похоже на еще один параметр d , который вы пытаетесь определить с помощью своего набора данных. Таким образом, первый вариант равен d , равному единице, если вы соответствуете линейной функции. Мы можем выбрать d равно двум, d равно трем, вплоть до d равно 10. Таким образом, мы хотели бы поместить этот дополнительный параметр, который я обозначаю d . И конкретно предположим, что вы хотите выбрать модель, то есть выбрать степень полинома, выбрать одну из этих 10 моделей. И подгоните эту модель, а также оцените, насколько хорошо ваша гипотеза была обобщена на новые примеры. Вот одна вещь, которую вы могли бы сделать. Что бы вы могли, сначала возьмите свою первую модель и минимизируйте ошибку обучения. И это даст вам некоторый вектор-параметр. И тогда вы могли бы взять свою вторую модель, квадратичную функцию и поместиться в соответствии с вашим набором тренировок, и это даст вам другое. Параметр θ . Чтобы различать эти различные векторы параметров, я собираюсь использовать надстрочный индекс, там, где $\theta^{\text{superscript}}$ просто означает параметры, которые я получаю, приспособивая эту модель к моим данным обучения. И верхний индекс θ_2 просто означает параметры, которые я получаю, устанавливая эту квадратичную функцию в мои данные обучения и так далее. Подбирая кубическую модель, я получаю скобки три до, ну, скажем, θ_{10} . И одна вещь, которую мы можем сделать, это взять эти параметры и посмотреть на ошибку теста. Поэтому я могу вычислить на своем тестовом наборе J тест одного, J теста θ_2 и т. Д.

3:47

J тест θ_3 и т. Д.

3:53

Поэтому я собираюсь взять каждую из своих гипотез с соответствующими параметрами и просто измерить производительность на тестовом наборе. Теперь одна вещь, которую я мог бы сделать, - это, чтобы выбрать одну из этих моделей, я мог бы увидеть, какая модель имеет самую низкую ошибку набора тестов. И давайте просто скажем для этого примера, что я решил выбрать полином пятого порядка. Таким образом, это покажется разумным. Но теперь предположим, что я хочу взять мою пятую гипотезу, эту, эту модель пятого порядка, и, скажем, хочу спросить, насколько хорошо эта модель обобщается?

4:27

Одна вещь, которую я мог бы сделать, это посмотреть, насколько хорошо моя полиномиальная гипотеза пятого порядка была выполнена на моем тестовом наборе. Но проблема в том, что это не будет справедливой оценкой того, насколько хорошо моя гипотеза обобщается. И причина в том, что мы сделали, так это то, что мы установили этот дополнительный параметр d , то есть эту степень полинома. И что соответствует этому параметру d , используя тестовый набор, а именно, мы выбрали значение d , которое дало нам наилучшую производительность в тестовом наборе. Итак, производительность моего вектора параметров θ_5 на тестовом наборе, вероятно, будет чрезмерно оптимистичной оценкой ошибки обобщения. Правильно, поэтому, потому что я поместил этот параметр d в свой тестовый набор, уже не справедливо оценивать мою гипотезу в этом тестовом наборе, потому что я поместил свои параметры в этот набор тестов, я выбрал степень d полинома, используя набор тестов. И поэтому моя гипотеза, скорее всего, будет лучше в этом тестовом наборе, чем на новых примерах, которые она не увидела раньше, и это то, что мне действительно интересно. Поэтому, чтобы повторить, на предыдущем слайде мы увидели, что если мы поместим какой-то набор параметров, вы знаете, скажем, θ_0 , θ_1 и т. Д., К некоторому набору тренировок, то производительность модели, установленной на обучающем наборе, не прогнозируя, насколько гипотеза будет обобщаться на новые примеры. Это связано с тем, что эти параметры соответствуют учебному набору, поэтому они, вероятно, хорошо справятся с набором тренировок, даже если параметры не подходят для других примеров. И в той процедуре, которую я только что описал в этой строке, мы просто сделали то же самое. И конкретно, что мы сделали, мы сопоставляем этот параметр d с тестовым набором. И, поместив параметр в тестовый набор, это означает, что эффективность гипотезы на этом наборе тестов может быть не справедливой оценкой того, насколько хорошо гипотеза, скорее всего, будет на примерах, которые мы не видели раньше. Чтобы решить эту проблему, в настройке выбора модели, если мы хотим оценить гипотезу, мы обычно это делаем. Учитывая набор данных, вместо того, чтобы просто раскалывать в набор тестовых тестов, то, что мы собираемся сделать, состоит в том, чтобы разделить его на три части. И первая часть будет называться учебным комплектом, как обычно.

6:50

Поэтому позвольте мне назвать эту первую часть тренировочным комплектом.

6:54

И вторая часть этих данных, я собираюсь назвать набор проверки перекрестных ссылок. [ЗВУК] Перекрестная проверка. И перекрестная проверка, как V-D. Иногда его также называют набором проверки, а не набором перекрестной проверки. И тогда потеря может заключаться в вызове обычного тестового набора. И симпатичное, довольно типичное соотношение, с помощью которого можно разделить эти вещи, - это отправить 60% ваших данных, ваш набор для обучения, возможно, 20% на ваш комплект перекрестной проверки и 20% на ваш тестовый набор. И эти цифры могут немного отличаться, но эта интеграция довольно типична. И поэтому наши обучающие наборы теперь будут всего лишь 60% данных, а наш набор перекрестной проверки или наш набор валидации будут иметь несколько примеров. Я собираюсь обозначить, что m subscript cv . Таким образом, это число примеров перекрестной проверки.

7:52

Следуя нашему раннему нотационному соглашению, я собираюсь использовать x_i cv sum u_i cv , чтобы обозначить пример перекрестного подтверждения i . И, наконец, у нас также есть тестовый набор, где наш тест под индексом m является числом тестовых примеров. Итак, теперь мы определили проверку валидации или перекрестные проверки и наборы тестов. Мы также можем определить ошибку обучения, ошибку перекрестной проверки и ошибку теста. Итак, вот моя ошибка в обучении, и я просто пишу это как J -индекс θ . Это почти то же самое. Это то же самое, что и J θ , о котором я писал до сих пор, это всего лишь ошибка набора учебных заданий, которую вы знаете, как измерение набора тренировок, а затем J -индекс cv моя ошибка перекрестной проверки, это в значительной степени то, что вы «Ожидайте, точно так же, как ошибка обучения, которую вы установили, измеряете ее на наборе данных перекрестной проверки, и вот моя тестовая тестовая ошибка такая же, как и раньше.

8:49

Поэтому, когда мы сталкиваемся с проблемой выбора модели, как это, то, что мы собираемся сделать, вместо использования набора тестов для выбора модели мы вместо этого будем использовать набор проверки или набор перекрестных проверок для выбора модель. Конкретно, мы сначала возьмем нашу первую гипотезу, возьмем эту первую модель и скажем, сведем к минимуму кросс-функцию, и это даст мне некоторый вектор-параметр для новой модели. И, как и прежде, я собираюсь поставить надстрочный индекс 1, просто чтобы обозначить, что это параметр для новой модели. Мы делаем то же самое для квадратичной модели. Получите некоторый вектор параметра θ_2 . Получите некоторый пара, вектор параметров θ_3 и т. Д., Вплоть до θ_{10} для полинома. И то, что я собираюсь сделать, - вместо того, чтобы тестировать эти гипотезы в тестовом наборе, я вместо этого буду тестировать их на наборе перекрестной проверки. И измерить J -индекс cv , чтобы увидеть, насколько хорошо каждая из этих гипотез делает на моем наборе перекрестной проверки.

9:53

И тогда я собираюсь выбрать гипотезу с самой низкой ошибкой проверки перекрестных ссылок. Итак, для этого примера, допустим, ради аргумента, что это мой полином 4-го порядка, который имел самую низкую ошибку перекрестной проверки. Поэтому в этом случае я собираюсь выбрать эту модель полинома четвертого порядка. И, наконец, что это означает, что этот параметр d , помните, что d была степенью полинома, правильно? Таким образом, d равно двум, d равно трем,

вплоть до d равно 10. То, что мы сделали, мы подберем этот параметр d , и мы скажем, что d равно четырем. И мы сделали это, используя набор кросс-валидации. И поэтому эта степень полинома, поэтому параметр больше не подходит для тестового набора, и поэтому мы не сохранили тестовый набор, и мы можем использовать набор тестов для измерения или оценить ошибку обобщения модель, которая была выбрана. К ним. Таким образом, это был выбор модели, и как вы можете взять свои данные, разделить ее на обучение, проверку и набор тестов. И используйте данные перекрестной проверки, чтобы выбрать модель и оценить ее на тестовом наборе.

10:59

В заключение, я должен сказать, что в машине. Как и в этой практике сегодня, не так много людей, которые будут делать то раннее, о чем я говорил, и сказали, что, знаете, это не такая хорошая идея, выбрать вашу модель с помощью этого набора тестов. И затем используя тот же набор тестов, чтобы сообщать об ошибке, как бы выбрав степень полинома в тестовом наборе, а затем сообщая об ошибке на тестовом наборе, как будто это была хорошая оценка ошибки обобщения. К сожалению, такого рода практика, многие люди делают это. Если у вас есть массовое массовое испытание, которое, возможно, не так уж и страшно, но многие практикующие, большинство практикующих, которые обучаются в машинах, склонны советовать против этого. И считается лучшей практикой иметь отдельные проверки валидации и тестовые наборы. Я просто предупреждал вас, чтобы иногда люди делали, как вы знаете, использовать те же данные для целей набора проверки и для набора тестов. Вам нужен набор для обучения и набор тестов, и это хорошо, это практика, хотя вы увидите, что некоторые люди это делают. Но, если это возможно, я бы рекомендовал не делать это самостоятельно.