Классификация и вероятности классов

Сеть выдала «безопасно» для двух мест, но одно закрашено густо, другое едва заметно. Что значит эта разница и стоит ли ей верить?

Робот по-прежнему получает две координаты и выбирает между «безопасно» и «опасно». В предыдущем сюжете мы добавили скрытые слои, чтобы сеть могла обвести безопасную область на кольце. Теперь разберём её последний ответ. Знак числа подсказывает класс, но что означает его величина? Почему две синие области бывают разной насыщенности?

Для первого опыта снова возьмём два облака и один нейрон: так проще проследить каждое вычисление. Скрытые слои меняют способ получения ответа, но вопрос остаётся тем же — как прочитать выход модели для конкретного места.

Ранее

Нейрон и многослойная сеть уже различают опасные и безопасные места по двум координатам. Теперь разберём величину выхода и вероятностное прочтение ответа.

← Полносвязные слои и многослойные сети

1. От числа к двум вероятностям

До сих пор выход $a = \tanh(s)$ лежал между -1 и 1. При $a > 0$ мы выбирали безопасный класс. Само число a нельзя читать как вероятность: оно может быть отрицательным. Введём явное правило преобразования:

$p(\text{безопасно}) = \frac{a+1}{2}$.

$p(\text{опасно}) = 1 - p(\text{безопасно})$.

При $a = -0{,}8$ получаем 10% для безопасного класса и 90% для опасного. При $a = 0$ доли одинаковы. При $a = 0{,}8$ соотношение обратное: 90% и 10%. Сумма всегда равна 100%, потому что классы взаимоисключающие и других вариантов в этой задаче нет.

Это новый способ чтения того же выхода: веса и карта от преобразования не изменяются. Порог $a > 0$ соответствует $p(\text{безопасно}) > 0{,}5$. На границе доли равны, и по нашей прежней договорённости выбираем «опасно».

Заметка

Здесь задана конкретная связь с Tanh. Цвет любой другой карты нельзя автоматически переводить в проценты по этому правилу. Вероятности модели ещё предстоит проверить на независимых примерах.

> Нейрон на карте вернул $a = -0{,}8$. Можно ли прочитать это число непосредственно как вероятность безопасного класса?
>
> - ○ Да: вероятность равна -80%
> - ○ Нет: нейрон обязательно сломан
> - ○ Да: достаточно убрать минус
> - ○ Нет: сначала нужно применить правило перевода выхода в вероятность

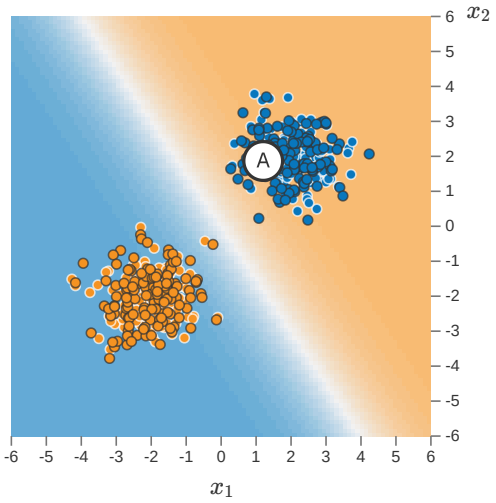
2. Одна точка, два возможных ответа

В тренажёре кнопки A, B, C выбирают **известные точки той же карты**. Маркер показывает выбранное место, рядом видны координаты, известный класс и ответ модели. Первоначальные веса

специально не подобраны: модель может уверенно ошибаться.

Попробуйте. Выберите A, установите оба веса равными 1 и оставьте смещение на 0. Посмотрите, какая доля досталась безопасному классу. Затем уменьшайте оба веса к нулю. Точка остаётся на месте, её известный класс не меняется, но распределение приближается к равным долям.

Что модель отвечает в выбранном месте



Два облака. Синий класс — безопасно, оранжевый — опасно.

Состояние опыта: 0 проходов. Скрытых слоёв нет.

Веса: -0,6; -0,4. Смещение: 0.

Для каждого пикселя считаем сумму: $s = w_1 x_1 + w_2 x_2 + b$

Точка A: (1,203; 1,866). Известный класс: безопасно. Прогноз: опасно.

$a \approx -0{,}899 \quad p = \frac{a+1}{2}$

Это вероятностная оценка модели. Её надёжность проверяют на новых данных.

Безопасно: 5,0 %
Опасно: 95,0 %

A, B и C — реальные обучающие точки на карте. Меняйте веса и сравнивайте класс, выход нейрона и две вероятности.

Интерактивный эксперимент — в электронной версии урока.

Теперь попробуйте отрицательные веса. Найдите настройку, при которой безопасная точка получила большой процент для опасного класса. Модель изменила мнение, но местность от этого не стала другой. Проверить ответ можно именно потому, что у выбранной точки уже есть метка.

Посмотрите и на точку B. Одна настройка весов используется сразу для всех мест. Нельзя независимо присвоить каждой точке понравившуюся вероятность: изменение одного веса пересчитывает

всю карту. Так выход отдельного примера связан с общим правилом модели.

При долях 55% и 45% выбирается тот же класс, что при 95% и 5%. Решение одинаково, но распределение различается. Поэтому одна метка скрывает часть информации. В следующем сюжете увидим, как функция потерь учитывает величину выхода, даже когда выбранный класс ещё не изменился.

> У места ровно один из двух классов. Модель дала безопасному классу 60%. Сколько процентов приходится на опасный?

Заметка

Вероятность относится к классу места целиком: 70% безопасности не означает, что точка «на 70% безопасна». Это оценка модели при данных входах.

3. Уверенная ошибка остаётся ошибкой

Большая вероятность — оценка модели, а не доказательство безопасности. Мы только что могли получить уверенно неверный ответ простым изменением весов. Ошибки возможны и после обучения, особенно на местах, непохожих на обучающие примеры.

Чтобы проверить смысл вероятностей, возьмём много новых мест, которым модель дала около 80% безопасности. Если примерно восемь из десяти таких мест действительно безопасны, оценки согласуются с частотой в этой группе. Такую согласованность называют **калибровкой**. По одной точке A её установить нельзя.

В нашем опыте мы вручную настраиваем параметры и рассматриваем учебные точки. Полученные проценты иллюстрируют вычисление, но этот опыт ещё не подтверждает качество вероятностей. Для проверки нужны отложенные данные, которые не использовали при подборе весов.

Правило выбора действия тоже задаём мы. В учебной классификации достаточно порога 0,5. Если ошибка имеет высокую цену, можно ввести другой порог или отправлять спорные случаи на дополнительную проверку. Это меняет решение по готовому ответу, а не сами веса сети.

> Модель дала опасному месту 99% вероятности безопасности. Что показывает этот пример?
>
> ○ Высокая вероятность не гарантирует правильный ответ
>
> ○ Нужно заменить 99% на 100%
>
> ○ Вероятности всегда бессмысленны
>
> ○ Все остальные ответы модели тоже ошибочны

4. Если классов станет больше

Допустим, робот теперь различает три типа поверхности: сухую землю, воду и болото. Для этого изменим постановку: у выходного слоя будет по одной исходной оценке на каждый из трёх классов.

Такие оценки называют **логитами**. Они могут быть отрицательными и не обязаны складываться в единицу.

Здесь используют **softmax**: он преобразует весь набор логитов в положительные доли с общей суммой 100%. Более высокой оценке достаётся большая доля. Оценки 2, 1, 0 превращаются примерно в 66,5%, 24,5%, 9%, а не в 200%, 100%, 0%.

Если поднять оценку воды, оставив две другие прежними, её доля вырастет, а остальные уменьшатся: все три связаны общим бюджетом. Для одной метки выбираем наибольшую долю. При равных максимальных оценках нужно отдельно договориться, как разрешать ничью.

Это расширение нашей задачи; тренажёр выше сохраняет два исходных класса и один выход. Для изображения, о котором говорили в сюжете о слоях, логика такая же: входов-пикселей много, а оценок на выходе столько, сколько предусмотрено классов. Если некоторого класса в списке нет, softmax сам его не добавит.

Следующий шаг

У нас есть координаты, прогноз и известный класс. Теперь зададим одно число, которое позволит сравнивать настройки весов: функцию потерь на этих же точках карты.

Выход Tanh переводим в две доли по явно заданному правилу; выбранная метка — отдельное решение.

Для нескольких классов softmax преобразует весь набор логитов в распределение.

Высокая вероятность может сопровождать ошибку; оценки проверяют на новых примерах.

Домашняя работа

1. Найдите для точки A две настройки с одним выбранным классом, но заметно разными вероятностями. Запишите веса и обе доли. Объясните, почему по одной метке эти настройки не различить.

2. Два взаимоисключающих класса карты имеют общий бюджет 100%. Безопасный получил 63%. Сколько получил опасный?

Источники

- Оценки классов и softmax (https://d2l.ai/chapter_linear-classification/softmax-regression.html)
- Оригинальный тренажёр взят из TensorFlow Playground (https://playground.tensorflow.org/). Авторы: Daniel Smilkov и Shan Carter.