Семинар 6.
ММП, осень 2012–2013
6 ноября

Темы семинара:

- Линейные методы классификации;
- Discriminant functions, Discriminative models, Generative models;
- Двухклассовая и многоклассовая логистическая регрессия;
- Неустойчивость логистической регрессии, введение регуляризации (априора).

1 Линейные методы классификации

Все рассматриваемые в лекциях методы классификации можно разделять на несколько разных подходов. Одни подходы делят решение задачи на два этапа: этап оценивания апостериорных распределений классов $p(Y|X)$ (известный как этап *вывода*, inference) и дальнейший этап использования этих оценок для классификации (принятия решения, decision). Другие подходы в явном виде строят отображение пространства объектов в пространство ответов $f\colon \mathbb{X} \to \mathbb{Y}$ (известное как *дискриминантная функция*, discriminant function), перескакивая этап оценивания распределений, и обычно эти методы не оперируют вероятностными интерпретациями.

- *Дискриминантная функция.* Поиск дискриминантной функции, которая напрямую отображает объекты x в пространство ответов $\mathbb{Y}$: $f\colon \mathbb{X} \to \mathbb{Y}$. Обучение в этом случае заключается в поиске конкретного вида дискриминантной функции (например, настройка параметров, если дискриминантная функция задана в виде параметрического семейства $f(x, \theta)$).

  Примеры: а) персептрон Розенблатта, б) классификация линейной регрессией и МНК, в) метод оптимальной разделяющей гиперплоскости (SVM).

- *Discriminative model, разделяющая модель.* Целью обучения является оценивания апостериорной вероятности классов $p(Y|X)$. Для этого мы представляем апостериорную вероятность в виде параметрического семейства функций $f(x, \theta)$ и обучение сводится к настройке параметра θ. Классификация осуществляется с помощью полученных оценок (с использованием формулы оптимального байесовского классификатора).

  Примеры: а) Логистическая регрессия, б) Пробит регрессия (probit regression).

- *Generative model, производящая модель*. Этот подход отличается от прошлого тем, что мы оцениваем апостериорную вероятность не напрямую, а в несколько этапов. Сначала оценим априорные вероятности классов $P(Y = k)$. Затем для каждого класса оценим распределение класса $p(X|Y)$. Наконец, используя формулу Байеса

  $$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)},$$

  мы получаем оценки апостериорных вероятностей.

  Подход называется *производящим*, поскольку в результате обучения мы получаем оценку распределения $p(X)$, из которого можем генерировать случайные наблюдения.

  Примеры: нормальный дискриминантный анализ, наивный байесовский классификатор.

На лекциях было приведено много аналогий, указывающих на сходство разделяющих и производящих подходов. Минимизация аппроксимации эмпирического риска может быть сведена к максимизации правдоподобия введением соответствующей функции потерь; а регуляризация, возникающая в вероятностных подходах в виде априорных распределений на параметрах модели, может быть получена добавлением нового слагаемого в функции потерь.

В случаях, когда разделяющая поверхность является линейной, метод классификации называется линейным.

На этом семинаре мы подробнее рассмотрим логистическую регрессию — пример разделяющего подхода.

2 Логистическая регрессия

Коротко напомним основную идею логистической регрессии в случае двух классов $\mathbb{Y} = \{+, -\}$.

Запишем апостериорную вероятность класса $+1$ в следующем виде:

$$p(y = +1|x) = \frac{p(x|y = +1)p(y = +1)}{p(x|y = +1)p(y = +1) + p(x|y = -1)p(y = -1)} =$$
$$= \frac{1}{1 + \exp(-a)} = \sigma(a),$$

где σ — сигмоидная функция и

$$a = \ln \frac{p(x|y = +1)p(y = +1)}{p(x|y = -1)p(y = -1)} = \frac{p(y = +1|x)}{p(y = -1|x)}. \tag{1}$$

В нормальном дискриминантном анализе мы видели, что если предположить нормальность распределений классов $p(X|Y)$ с одинаковой ковариационной матрицей, то величина a в (1) примет линейный вид $a = \boldsymbol{w}^\top \boldsymbol{x} + w_0$ (вследствие сокращения квадратичных членов из-за совпадающих ковариационных матриц). Поскольку разделяющая поверхность оптимального байесовского классификатора соответствует линиям уровня апостериорной вероятности класса (не обязательно $p(Y|X) = 0.5$, так как штрафы λ_k могут быть разными), в этом случае разделяющая поверхность будет линейна и иметь вид $\boldsymbol{w}^\top \boldsymbol{x} + w_0 = const$. Для её поиска мы строили выборочные оценки ковариационной матрицы и средних, а также априорные вероятности классов, и затем уже получали вид разделяющей поверхности. Это пример производящего подхода.

Более того, на лекции был установлен интересный факт: если распределения классов принадлежат к экспоненциальному семейству и их параметры удовлетворяют определенным условиям, то величина (1) снова принимает линейный вид. Этот факт может быть использован в качестве обоснования логистической регрессии.

Логистическая регрессия предполагает, что выражение (1) действительно представимо в линейном виде, а значит апостериорная вероятность класса записывается в следующем виде:

$$p(Y = +1|X) = \sigma(a) = \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle). \tag{2}$$

Мы предположили, что в признаках объекта содержится константный признак, поэтому свободный член w_0 вошел в скалярное произведение. Также легко заметить, что $p(y|x) = \sigma(y\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$.

Теперь в отличие от производящего подхода мы можем напрямую настраивать параметр $\boldsymbol{w}$ с помощью максимизации правдоподобия:

$$L(w, X^\ell) = \sum_{i=1}^{\ell} \ln p(x_i, y_i) \to \max_{\boldsymbol{w}}.$$

Учитывая, что $p(x, y) = p(y|x)p(x)$ и второй множитель не зависит от $\boldsymbol{w}$, мы приходим к эквивалентной задаче

$$E(\boldsymbol{w}) = \sum_{i=1}^{\ell} \ln p(y_i|x_i) = \sum_{i=1}^{\ell} \ln \sigma(y_i\langle \boldsymbol{w}, x_i \rangle) \to \max_{\boldsymbol{w}}, \tag{3}$$

или, что то же самое,

$$\sum_{i=1}^{\ell} \ln\bigl(1 + \exp\left(-y_i\langle \boldsymbol{w}, x_i \rangle\right)\bigr) \to \min_{\boldsymbol{w}}. \tag{4}$$

В случае совпадающих штрафов $\lambda_1 = \lambda_2$ разделяющая поверхность будет иметь вид $P(Y|X) = 0.5$. В нашем случае апостериорная вероятность выражается с помощью сигмоидной *функции активации* и точка $p(y|x) = \sigma(y\langle \boldsymbol{w}, x \rangle) = 0.5$ соответствует $y\langle \boldsymbol{w}, x \rangle = 0$. Таким образом в этом случае классификация соответствует правилу $a(x) = \operatorname{sgn}\langle \boldsymbol{w}, x \rangle$, то есть точки пространства разделяются на классы гиперплоскостью с нормалью $\boldsymbol{w}$.

Пользуясь правилом $\sigma(a)'_a = \sigma(z)\sigma(-a)$ можно легко найти градиент функционала в задаче (3):

$$\nabla E(\boldsymbol{w}) = \sum_{i=1}^{\ell} x_i y_i \sigma(-y_i\langle \boldsymbol{w}, x_i \rangle)$$

и организовать итерации метода стохастического градиента, пользуясь j-м слагаемым в представлении градиента при вытягивании j-го объекта. Более надежным методом решения этой оптимизационной задачи является применение метода Ньютона-Рафсона — метода второго порядка. Об этом будет рассказано в одной из следующих глав курса.

Задача. *Опишите вид разделяющей поверхности логистической регрессии. Вспомните понятие отступа объекта относительно линейного классификатора и убедитесь, что в случае $\lambda_1 = \lambda_2$ задача (4) эквивалентна минимизации сглаженного эмпирического риска.*

Заметим, что логистическая регрессия требует настройки $M+1$ параметров вектора $\boldsymbol{w}$, где M — размерность исходного пространства. В то же время ЛДФ настраивает $M|\mathbb{Y}|+M(M+1)/2+|\mathbb{Y}|-1$ параметров. Это следствие разницы в производящих и разделяющих подходах.

Задача. *Сравните ЛДФ и логистическую с точки зрения предположений, налагаемых на маргинальное распределение $p(X)$.*

2.1 Многоклассовая логистическая регрессия.

В случае многих классов $\mathbb{Y} = \{1, \dots, K\}$ метод меняется не сильно. На этот раз мы предполагаем линейность следующих выражений:

$$
\begin{aligned}
&\ln \frac{p(Y=1|X=x)}{p(Y=K|X=x)} = \langle \boldsymbol{w}_1, x\rangle;\\
&\ln \frac{p(Y=2|X=x)}{p(Y=K|X=x)} = \langle \boldsymbol{w}_2, x\rangle;\\
&\dots\\
&\ln \frac{p(Y=K-1|X=x)}{p(Y=K|X=x)} = \langle \boldsymbol{w}_{K-1}, x\rangle.
\end{aligned}
$$

Задача: *получите выражения для $p(y=k|x)$ для $k = 1, \dots, K$.*

$$
\begin{aligned}
p(Y=k|X=x) &= \frac{\exp(\langle \boldsymbol{w}_k, x\rangle)}{1+\sum_{i=1}^{K-1}\exp(\langle \boldsymbol{w}_i, x\rangle)},\ k = 1\ \dots, K-1;\\
p(Y=k|X=x) &= \frac{1}{1+\sum_{i=1}^{K-1}\exp(\langle \boldsymbol{w}_i, x\rangle)},\ k = K.
\end{aligned}
$$

2.2 Численная неустойчивость решения логистической регрессии.

Два множества точек A и B *линейно разделимы*, если существует вектор w и число w_0 такие, что

$$
\begin{cases}
\langle a, w\rangle + w_0 > 0, & \text{для всех } a \in A;\\
\langle b, w\rangle + w_0 < 0, & \text{для всех } b \in B.
\end{cases}
$$

Задача. *Исследуйте ОМП оценку для вектора $\boldsymbol{w}$ логистической регрессии в случае двух классов и линейно разделимой выборки.*

Оказывается, что в случае линейной разделимости обучающей выборки решение задачи (3) будет неустойчивым. В этом случае существует вектор $\boldsymbol{w}$, который удовлетворяет условиям из определения линейной разделимости выборки, то есть для него $\langle \boldsymbol{w}, x_i \rangle > 0$, если $y_i = +1$ и $\langle \boldsymbol{w}, x_i \rangle < 0$ если $y_i = -1$. Следовательно, зафиксировав любой такой вектор $\boldsymbol{w}$ и затем бесконечно увеличивая его норму, мы будем бесконечно уменьшать наш функционал. Апостериорные распределения при этом выродятся в пороговую функцию. Кроме того, если обучающая выборка строго разделима, то существует континуум гиперплоскостей, безошибочно разделяющих обучающую выборку. К какой из них сойдется метод стохастического градиента зависит от инициализации.

Чтобы избежать описанных проблем целесообразно вводить в оптимизационную задачу (3) регуляризатор. На вероятностном языке это означает введение *априорного* распределения $p(\boldsymbol{w})$ на векторе $\boldsymbol{w}$ и максимизацию не правдоподобия обучающей выборки, а апостериорной вероятности $p(\boldsymbol{w}|X^\ell)$ вектора $\boldsymbol{w}$, что эквивалентно следующему:

$$\ln p(\boldsymbol{w}|X^\ell) = \ln p(X^\ell|\boldsymbol{w}) + \ln p(\boldsymbol{w}) - \ln p(X^\ell) \to \max_{\boldsymbol{w}};$$
$$\ln p(X^\ell|\boldsymbol{w}) + \ln p(\boldsymbol{w}) \to \max_{\boldsymbol{w}}.$$

Вводя различные априорные распределения можно получать различные регуляризаторы. Нормальный априор ведет к квадратичной регуляризации, а лапласовский — к L_1 регуляризации.