1 Классификация

1.1 Общий подход к классификации через апостериорные вероятности

Общая подход к классификации: строятся классифицирующие функции f_i, такие что классификация проводится так: индивид с признаками x относится к группе с максимальным значением на нем классифицирующей функции: $\arg\max_i \mathrm{f}_i(x)$.

Откуда берутся эти классифицирующие функции? Естественная идея взять в качестве f_i вероятность (ее оценку) принадлежности к i-му классу. Пусть ξ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{P}_i$ и имеет плотность $p_i(x)$. Тогда было бы логично взять $f_i = p_i$. Для практического применения надо было бы оценить плотности, либо непараметрически (например, по числу точек, попавших в дельта-окрестность — типа метода ближайших соседей), либо параметрически (если известно, что распределение нормальное, тогда просто оцениваем векторы средних и ковариационные матрицы).

Более сложный подход — через апостериорные вероятности. Если у нас есть априорное знание вероятности того, что индивид из того или иного класса, то мы можем его учесть. Введем понятие класса $C_i = \{\xi = A_i\}$. Чтобы классифицировать наблюдение x, необходимо найти

$$\arg\max \mathrm{P}\left(\xi \in A_i | \eta = x\right) = \arg\max \mathrm{P}\left(C_i | x\right).$$

Пусть известны априорные вероятности принадлежности нового наблюдения к i-му классу $\pi_i = \mathrm{P}\left(C_i\right)$. Тогда апостериорные вероятности по формуле Байеса будут иметь вид

$$\mathrm{P}\left(C_i | x\right) = \frac{\mathrm{P}\left(x | C_i\right)\pi_i}{\sum_{j=1}^k \mathrm{P}\left(x | C_j\right)\pi_j}.$$

Поэтому в качестве классифицирующих функций берут

$$f_i\left(x\right) = \frac{p_i(x)\pi_i}{\sum_{j=1}^k p_j(x)\pi_j}.$$

Так как знаменатель у всех f_i одинаковый, его можно отбросить, и итоговые классифицирующие функции будут выглядеть как $f_i\left(x\right) = \mathrm{P}\left(x | C_i\right)\pi_i = p_i(x)\pi_i$.

Как выбрать априорные вероятности?

1. Равномерно, $\forall i \in 1 : k\ \pi_i = 1\ /\ k$.
2. По соотношениям в обучающей выборке: $\pi_i = n_i\ /\ \sum_{j=1}^k n_j$.
3. На основе другой дополнительной информации о данных (результаты предыдущих исследований, etc.)

Свойство. *Построенный метод классификации* $\mathrm{predict}(x) = \arg\max_i \pi_i p_i(x)$ *минимизирует среднюю апостериорную ошибку:*

$$\sum_{i=1}^k \pi_i P(\mathrm{predict}(x) != i \mid C_i).$$

Видно, что можно с помощью априорных вероятностей формально задавать важность ошибочных классификаций для разных классов.

1.2 Линейный и квадратичный дискриминантный анализ для классификации

1.2.1 LDA

Модель: ξ — дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{N}(\mu_i, \Sigma)$. Тогда плотность в точке x

$$p_i(x) = p(x|\xi = A_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma^{-1}(x-\mu_i)\right),$$

и классифицирующая функция $f_i(x) = \pi_i p(x|\xi = A_i)$, где π_i — априорная вероятность наблюдения попасть в i-ю группу. Для упрощения вычислений можно переписать классифицирующую функцию через возрастающее монотонное преобразование как

$$g_i(x) = \log f_i(x) = \log \pi_i - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma^{-1}(x-\mu_i).$$

Сократив часть, не зависящую от номера класса, получаем линейные классифицирующие функции

$$h_i(x) = -\frac{1}{2}\mu_i^{\mathrm{T}}\Sigma^{-1}\mu_i + \mu_i^{\mathrm{T}}\Sigma^{-1}x + \log \pi_i.$$

1.2.2 QDA

Модель: ξ — дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\mathcal{P}(\eta \mid \xi = A_i) = \mathcal{N}(\mu_i, \Sigma_i)$. Тогда плотность в точке x

$$p(x|\xi = A_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i)\right),$$

и классифицирующая функция $f_i(x) = \pi_i p(x|\xi = A_i)$. Применяем возрастающее монотонное преобразование и оставляем в классифицирующей функции только члены, отличающиеся в разных группах:

$$g_i(x) = \log f_i(x) = \log \pi_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x-\mu_i),$$

получаем квадратично зависящую от x классифицирующую функцию.

1.3 Классификация в случае двух классов

Если всего два класса, то можно построить границу между классами, приравняв классифицирующие функции.

1.3.1 LDA

Приравняв $h_1(x) = h_2(x)$, получим разделяющую гиперплоскость. Разделяющая два класса гиперплоскость имеет вид

$$\{\mathbf{x} : h_1(\mathbf{x}) = h_2(\mathbf{x})\} = \\ = \{\mathbf{x} : -\frac{1}{2}(\mu_1-\mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mu_1+\mu_2) + (\mu_1-\mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} + \log(\pi_1/\pi_2) = 0\}.$$

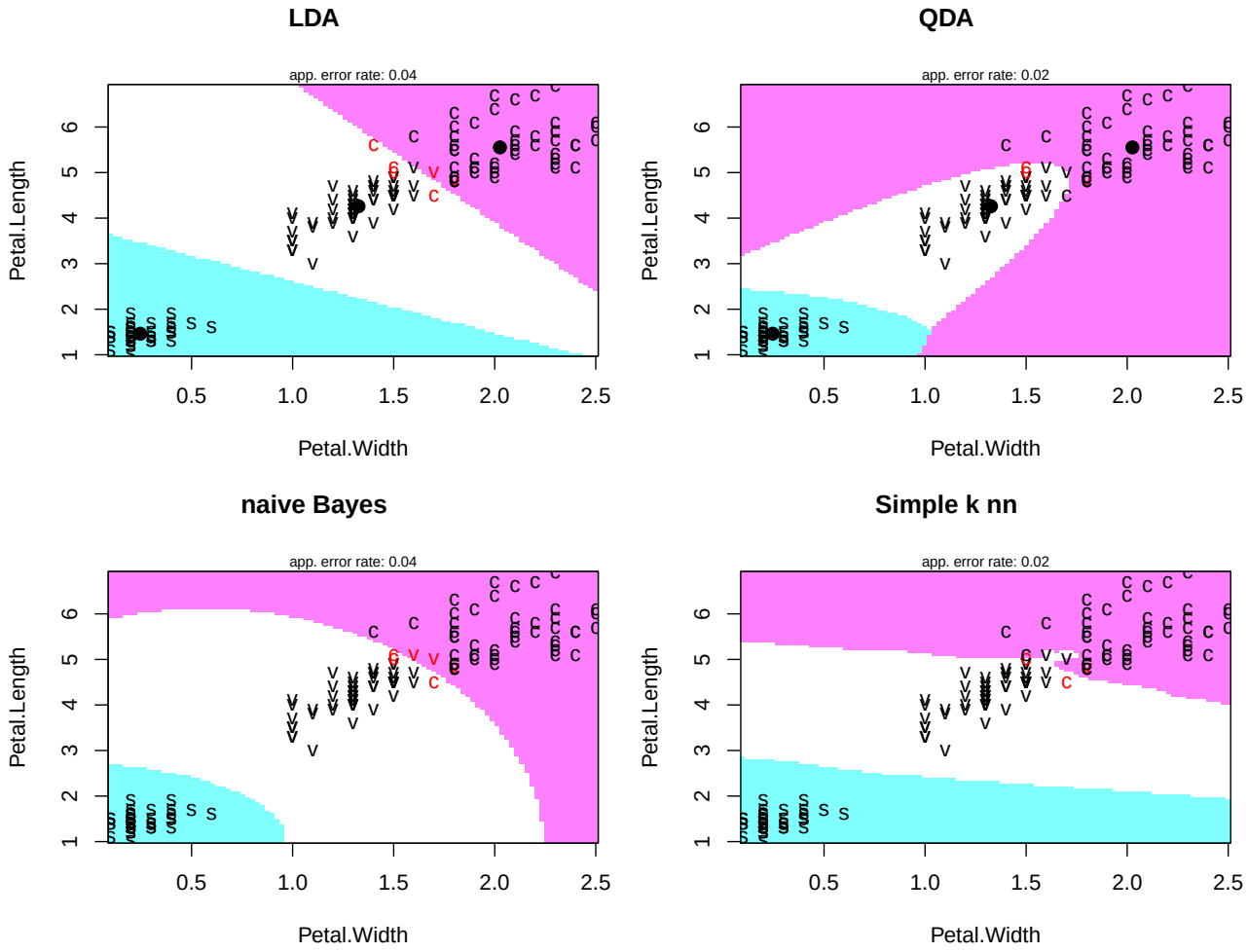
От соотношения между априорными вероятностями зависит положение границы относительно классов (к какому она ближе). Видно, что априорные вероятности влияют только на сдвиг разделяющей гиперплоскости.

Заметим, что классификацию можно записать как сравнение $-\frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mu_1 + \mu_2) + (\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x}$ с некоторым порогом $(-\log(\pi_1/\pi_2))$, который зависит от априорных вероятностей (или весов ошибок для разных классов, смотря как на это смотреть).

1.3.2 QDA

В данном случае, разделяющая поверхность имеет вид квадратичной поверхности, может состоять из двух гиперболоидом, может иметь форму эллипса.

1.3.3 Картинки



Здесь мы обсуждали число параметров в моделях, возможный overfitting (переподгонку). Использовали слова — обобщающая способность алгоритма.

1.4 Качество классификации

1.4.1 Ошибки класификации

Качество классификации измеряется ошибками классификации (доля неправильно классифицированных объектов). n_{ij} — число объектов из класса i, отнесенных к классу j. В соответствующей матрице классификации на диагонали стоят правильно классифицированные объекты, вне диагонали — ошибки.

На самом деле, нельзя проверять качество предсказания на тех данных, на которых это предсказание строилось. Поэтому используют кросс-валидацию (скользящий контроль). Например, каждое наблюдение по очереди исключается из выборки, классифицирующее правило строится без него и с помощью этого правила индивид классифицируется. Строится аналогичная таблица из n_{ij}. В ней ошибок будет, вообще говоря, больше.

Здесь обсуждали, что имеет смысл смотреть на ошибки без кросс-валидации и с ней. Если разница существенная, то это говорит о переподгонке используемой модели. Вероятно, она не очень хорошая; например, слишком много параметров.

Замечание. Нельзя путать классификацию и различие групп. Группы могут значимо различаться, классификация может быть при этом бессмысленной (ошибок чуть меньше 50%).

1.4.2 ROC и AUC

wikipedia ROC-кривая (англ. receiver operating characteristic, рабочая характеристика приёмника) — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, (англ. true positive rate, TPR, называемой чувствительностью алгоритма классификации) и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак (англ. false positive rate, FPR, величина 1-FPR называется специфичностью алгоритма классификации) при варьировании порога решающего правила.

Также известна как кривая ошибок. Анализ классификаций с применением ROC-кривых называется ROC-анализом.

Количественную интерпретацию ROC даёт показатель AUC (англ. area under ROC curve, площадь под ROC-кривой) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). Значение менее 0,5 говорит, что классификатор действует с точностью до наоборот: если положительные назвать отрицательными и наоборот, классификатор будет работать лучше.

мои комментарии Если кто-то хорошо представляет себе, как выглядит график зависимости мощности от ошибки первого рода, то это именно такой график. Меняется уровень значимости (как порог отвергнуть - не отвергнуть) и по оси x откладывается ошибка первого рода, она же false positive rate FP/(TN+FP), а по оси y откладывается мощность, она же true positive rate TP/(TP+FN)(слово positive означает, что нулевая гипотеза отвергнута в пользу второй, альтернативной, гипотезы, а в случае классификации, что элемент классифицируется как относящийся ко второму классу).

Таким образом, меняем порог/параметр для метода классификации (пример параметра — априорная вероятность π_1) и по оси x откладываем долю неправильно классифицированных элементов из первого класса $(n_{12}/(n_{11}+n_{12}))$, а по оси y — долю правильно классифицированных элементов из второго класса $(n_{22}/(n_{22}+n_{21}))$.

Пусть классы имеют вид 4,6,8,10,12 первый и 1,3,5,7 второй. Опишем ROC-кривую. Пусть к первому классу мы относим, если число больше порога γ. Для $\gamma < 1$ мы находимся в точке $(0,0)$. При $1 < \gamma < 3$ мы перескакиваем в точку $(0, 0.25)$. При $3 < \gamma < 4$ мы перескакиваем в точку $(0, 0.5)$. При $4 < \gamma < 5$ мы перескакиваем в точку $(0.2, 0.5)$. При $5 < \gamma < 6$ мы перескакиваем в точку $(0.2, 0.75)$. При $6 < \gamma < 7$ мы перескакиваем в точку $(0.4, 0.75)$. При $7 < \gamma < 8$ мы перескакиваем в точку $(0.4, 1)$. Дальше мы при $x = 1$ последовательно перескакиваем по y в 0.6, 0.8 и при $\gamma > 12$ попадаем в точку (1,1).