1 Матричная алгебра

1. Пусть $\mathbf{C} \in \mathbb{R}^{p\times p}$ — симметричная неотрицательно-определенная матрица, $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ — ее собственные числа. Напомню, что след квадратной матрицы (trace) — это сумма значений на диагонали. $|\cdot|$ здесь обозначает определитель.

   Тогда $\text{trace}(\mathbf{C}) = \sum_{i=1}^{p} \lambda_i$, $|\mathbf{C}| = \prod_{i=1}^{p} \lambda_i$, $|\mathbf{I}+\mathbf{C}| = \prod_{i=1}^{p}(1+\lambda_i)$, $1/|\mathbf{I}+\mathbf{C}| = \prod_{i=1}^{p} 1/(1+\lambda_i)$.

2. Пусть $\mathbf{A} \in \mathbb{R}^{p\times p}$ — симметричная неотрицательно-определенная матрица ранга r, Пусть $\mathbf{B} \in \mathbb{R}^{p\times p}$ — симметричная положительно-определенная матрица. Тогда матрица $\mathbf{B}^{-1}\mathbf{A}$ (несимметричная, вообще говоря) имеет p неотрицательных собственных чисел $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. Число ненулевых собственных чисел равно рангу r матрицы $\mathbf{A}$. Соответствующие собственные вектора образуют базис, который можно выбрать так, чтобы $\mathbf{U}^{\mathrm{T}}\mathbf{B}\mathbf{U} = \mathbf{I}$, где $\mathbf{U}$ — матрица, составленная из собственных векторов (говорят, что собственные вектора ортонормированы относительно матрицы $\mathbf{B}$).

   Так как $\mathbf{B}^{-1}\mathbf{A}U = \lambda U$ равносильно $\mathbf{A}U = \lambda\mathbf{B}U$, то эту задачу называют обобщенной задачей на собственные числа и собственные вектора (вместо единичной матрицы стоит симметричная положительно определенная матрица $\mathbf{B}$).

   Для обобщенной задачи на собственные значения справедливы те же свойства оптимальности, что и для обычной, только с ортогональностью относительно матрицы $\mathbf{B}$. А именно, $\sup_U U^{\mathrm{T}}\mathbf{A}U/U^{\mathrm{T}}\mathbf{B}Z = \sup_{U:U^{\mathrm{T}}\mathbf{B}U=1} U^{\mathrm{T}}\mathbf{A}U$ равен максимальному собственному числу λ_1 и достигается на $Z = U_1$; $\sup_{U:U^{\mathrm{T}}\mathbf{B}U_1=0} U^{\mathrm{T}}\mathbf{A}U/U^{\mathrm{T}}\mathbf{B}U$ равен λ_2 и достигается на $U = U_2$. И т.д.

3. В условиях предыдущего пункта, $|\mathbf{B}|/|\mathbf{B}+\mathbf{A}| = 1/|\mathbf{I}+\mathbf{B}^{-1}\mathbf{A}| = \prod_{i=1}^{r} 1/(1+\lambda_i)$.

   Следствие. Пусть $\mathbf{A}$ имеет распределение Уишарта $W_p(\mathbf{I}, \nu_A)$, $\mathbf{B}$ имеет распределение Уишарта $W_p(\mathbf{I}, \nu_B)$, $\nu_B \geq p$. $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ — собственные числа $\mathbf{B}^{-1}\mathbf{A}$. Тогда $\prod_{i=1}^{s} 1/(1+\lambda_i)$, где $s = \min(p, \nu_A)$, имеет распределение Лямбда Уилкса $\Lambda_p(\nu_A, \nu_B)$.

Questions

1. В обычной, не обобщенной, задаче на собственные числа, если ее привязать к SVD, получим случай $\mathbf{A} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{B} = \mathbf{I}$. У нас есть утверждение про эффективность SVD (про главные направления). Почему задача $\sup_{U:U^{\mathrm{T}}\mathbf{B}U=1} U^{\mathrm{T}}\mathbf{A}U$ и ответ на нее в виде $U = U_1$ в этом случае соответствует теореме про оптимальность SVD (что норма вектора Z_1 из первых главных компонент максимальна)?

2. Объясните, почему верен пункт 3 из теории выше и следствие из него.

2 Множественная регрессия и one-way ANOVA

Ниже будем считать, что все случайные величины имеют нормальное распределение.

1. Multiple regression: η, $\zeta = (\zeta_1, \ldots, \zeta_k)^{\mathrm{T}}$ — количественные признаки. Пусть $\hat{\eta}$ — наилучшее линейное приближение по МНК. В случае нормальной модели, это $\hat{\eta} = \mathbb{E}(\eta|\zeta_1, \ldots, \zeta_k)$.

Тогда получаем разложение дисперсии

$$\mathbb{E}(\eta - \mathbb{E}\eta)^2 = \mathbb{E}(\hat{\eta} - \mathbb{E}\eta)^2 + \mathbb{E}(\eta - \hat{\eta})^2. \tag{1}$$

Разложение суммы квадратов имеет вид

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_i(\hat{y}_i - \bar{y})^2 + \sum_i(y_i - \hat{y}_i)^2. \tag{2}$$

Получаем: SStotal = SSregression + SSerror и соотношение для степеней свободы $(n-1) = (k) + (n-k-1)$. Здесь имеется в виду, что каждая сумма, деленная на дисперсию остатков регрессии $\eta - \hat{\eta}$, имеет в нормальной модели распределение хи-квадрат с соответствующим числом степеней свободы.

Гипотеза о незначимости регрессии $H_0 : \mathbb{E}(\hat{\eta} - \mathbb{E}\eta)^2 = 0$ (наилучшее предсказание — просто среднее). Она проверяется с помощью статистики критерия: $t = SSregr/(k)/(SSerror/(n-k-1)) \sim F(k, n-k-1)$. Если гипотеза отвергается, то регрессия значима.

2. One-way ANOVA: η — количественный признак, ξ — одномерный качественный признак, принимающий значения $A_1, \ldots, A_k$.

Тогда получаем разложение дисперсии

$$\mathbb{E}(\eta - \mathbb{E}\eta)^2 = \mathbb{E}(\mathbb{E}(\eta|\xi) - \mathbb{E}\eta)^2 + \mathbb{E}(\eta - \mathbb{E}(\eta|\xi))^2. \tag{3}$$

Разложение суммы квадратов имеет вид

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2. \tag{4}$$

Получаем: SStotal = SSbetween + SSwithin и соотношение для степеней свободы $(n-1) = (k-1) + (n-k)$.

Гипотеза $H_0 : \mathbb{E}(\mathbb{E}(\eta|\xi) - \mathbb{E}\eta)^2 = 0$ (о равенстве средних в группах, соответствующих фиксированным значениям ξ). Если $\hat{\eta} = \mathbb{E}(\eta|\xi)$ не зависит от ξ, то $\hat{\eta} = \mathbb{E}\eta$.
Гипотеза проверяется с помощью статистики критерия: $t = SSbetween/(k-1)/(SSwithin/(n-k)) \sim F(k-1, n-k)$. Это же называется "провести ANOVA".

3. Введем фиктивные (dummy) переменные $\zeta_1, \ldots, \zeta_{k-1}$ на основе качественного признака ξ: $\zeta_i = 1$, если $\xi = A_i$; иначе $\zeta_i = 0$.

Тогда разложения дисперсии для ANOVA превращается в разложение дисперсии для линейной регрессии η на $\zeta_1, \ldots, \zeta_{k-1}$.

Корреляционное отношение η к ξ равно множественному коэффициенту корреляции между η и $\zeta_1, \ldots, \zeta_{k-1}$ (построение наилучшего функционального предсказания от ξ эквивалентно построению наилучшего линейного предсказания от $\zeta_1, \ldots, \zeta_{k-1}$).

4. Введем единые обозначения, SST = SSH + SSE (H — hypothesis, E — error), ν_H и ν_E соответствующее число степеней свободы.

Тогда критерий выглядит как $F = (SSH/\nu_H)/(SSE/\nu_E) \sim F(\nu_H, \nu_E)$. Критическая область — справа. При этом $R^2 = SSH/SST$ — это коэффициент детерминации, или оценка множественного коэффициента корреляции, или оценка корреляционного отношения.

5. Обозначим $\lambda_1 = SSH/SSE$.

Тогда корреляция $r^2 = \lambda_1/(1+\lambda_1)$. Чем больше r^2, тем больше значимость ANOVA/регрессия.

$F = \lambda_1 \cdot (\nu_E/\nu_H) \sim F(\nu_H, \nu_E)$. Чем больше λ (и F), тем больше значимость ANOVA/регрессия.

$\Lambda = 1/(1+\lambda_1) = 1 - r^2 \sim \Lambda_1(\nu_H, \nu_E)$. Чем **меньше** Λ, тем больше значимость ANOVA/регрессия.

Questions

Сначала комментарии. Линейная регрессия — это наилучшее приближение по методу наименьших квадратов одной случайной величины линейной функцией от набора случайный величин. Условное математическое ожидание — это наилучшее приближение по методу наименьших квадратов одной случайной величины произвольной функцией от набора случайный величин. Если все величины имеют совместное нормальное распределение, то известно, что наилучшее линейное предсказание задаётся линейной функцией (т.е. просто регрессия и линейная регрессия совпадают.) Условное математическое ожидание — это функциональная регрессия, или просто регрессия. Если $\widehat{\eta}(\xi_1, \ldots, \xi_k)$ — регрессия, то ошибка предсказания по ней дает минимально возможную среднеквадратическую ошибку.

Когда мы ищем ближайшую точку в линейном подпространстве (если два элемента ему принадлежат, то и любая их линейная комбинация тоже), то мы как бы опускаем перпендикуляр на это подпространство. Поэтому получаем, что это наилучшее приближение (ближайшая точка) перпендикулярно разнице между η и ее приближением.

Если рассмотреть линейное пространство случайных величин с нулевым мат.ожиданием, то $||\zeta||^2 = \mathbb{E}\zeta^2$. В пространстве векторов, норма обычная (или с делением на n, если мера вероятностная). Расстояние, как обычно, определяется как норма разности.

По обозначения в ANOVA. Можно рассматривать данные как один признак количественный (например, рост), а второй качественный (например, из какого материала сделаны кастрюли, в которых обычно варится еда — когда-то говорили, что сейчас все стали выше ростом, так как ели алюминий :)). Тогда y_{ij} — это просто перенумерация значений из первого признака, если мы перенумеруем всех так, что первый индекс — тип кастрюли, а второй — номер по порядку человека, который пользовался i-й кастрюлей (всего таких людей n_i, их средний рост $\bar{y}_i$). Поэтому сумма квадратов центрированных значений y_{ij} — это просто норма вектора из ростов. Вообще, норма вектора — это сумма по одному индексу, а здесь по двум, потому что просто так перенумеровали.

1. Почему все эти разложения (1), (2), (3), (3) можно назвать теоремой Пифагора?

2. Почему гипотеза о незначимости линейной регрессии (что равносильно тому, что все коэффициенты кроме b_0 равны нулю) имеет вид $H_0 : \mathbb{E}(\hat{\eta} - \mathbb{E}\eta)^2 = 0$? Почему предлагаемая статистика критерия измеряет то, насколько данные отличаются от гипотезы, где критическая область?

3. Почему гипотеза в ANOVA эквивалентна $H_0 : \mathbb{E}(\mathbb{E}(\eta|\xi) - \mathbb{E}\eta)^2 = 0$? Почему предлагаемая статистика критерия измеряет то, насколько данные отличаются от гипотезы, где критическая область? (Напомню: здесь ξ — одномерный признак с k значениями).

4. Почему если мы превратим признак ξ с k значениями в dummy признаки из нулей и единиц, то наилучшее функциональное приближение совпадет с наилучшим линейным приближением от dummy признаков? (Подсказка: функция ϕ, заданная на конечном числе значений $A_1, \ldots, A_k$, определяется принимаемыми k значениями $\phi_1 = \phi(A_1)$, ..., $\phi_k = \phi(A_k)$.)

5. Выпишите, чему равны SSH, SSE, ν_H, ν_E отдельно для линейной регрессии и отдельно для дисперсионного анализа.

6. Как видно из текста, с помощью статистики F, определяемой однотипно для линейной регрессии и дисперсионного анализа, можно проверять гипотезу о незначимости соответствующего вида анализа. Если гипотеза отвергается, то, значит, анализ значим (и его имеет смысл проводить). При этом для статистики F критическая область справа. Так как r^2, Λ выражаются через те же величины, их тоже можно использовать для проверки той же гипотезы. Где должна быть критическая область в каждом из случаев?