12. Übungsblatt zur Vorlesung
Statistische Methoden der Datenanalyse
Abgabe: 30.01.2016, 23:59 Uhr

WS 2017/2018
Prof. W. Rhode

Zeit	Raum	Abgabe im Moodle; Mails mit Betreff: [SMD1718]
Di. 10-12	CP-03-150	philipp2.hoffmann@udo.edu und jan.soedingrekso@udo.edu
Di. 16-18	P1-02-110	felix.neubuerger@udo.edu und tobias.hoinka@udo.edu
Di. 16-18	CP-03-150	simone.mender@udo.edu und maximilian.meier@udo.edu

Aufgabe 36: *Entfaltung in zwei Intervallen* — 6 P.

Betrachten Sie ein Experiment, in dem 2 Typen von Ereignissen gezählt werden. Die Wahrscheinlichkeit ein Ereignis dem falschen Typ zuzuordnen sei ϵ und die Wahrscheinlichkeit für den richtigen Typ $1-\epsilon$. Die Wahrscheinlichkeit der falschen Zuordnung ist also für $1 \to 2$ die gleiche wie für $2 \to 1$. Des weiteren gehen bei dem Experiment leider 20% der Messungen verloren.

Die gemessen Ereigniszahlen $\mathbf{g} = (g_1, g_2)^T$ sind Poisson-verteilt und unkorreliert.

- **a)** Stellen Sie die Antwortmatrix $\mathbf{A}$ auf und den Zusammenhang zwischen $\mathbf{A}$, $\mathbf{g}$ und der wahren Ereigniszahl $\mathbf{f} = (f_1, f_2)^T$.
- **b)** Berechnen Sie $\mathbf{f}$ als Funktion von $\mathbf{g}$ und ϵ.
- **c)** Berechnen Sie die Kovarianzmatrix $\mathbf{V}[\mathbf{f}]$ als Funktion von $\mathbf{g}$ und ϵ.
- **d)** Berechnen Sie $\mathbf{f}$, $\mathbf{V}[\mathbf{f}]$, die Fehler von f_1 und f_2 und den Korrelationskoeffizienten für $g_1 = 200$, $g_2 = 169$ und $\epsilon = 0,1$.
- **e)** Wie **d)**, aber mit $\epsilon = 0,4$. Was hat sich geändert und was bedeutet dies?
- **f)** Was passiert bei $\epsilon = 0,5$?

Aufgabe 37: *Entfaltung mit quadratischen Matrizen* — 7 P.

In dieser Aufgabe sollen die gezeigten Plots aus der Vorlesung nachgebaut werden, um die Vorgänge bei der Entfaltung besser zu verstehen.

- **a)** Die quadratische $n \times n$ Antwortmatrix A wird hier nicht aus MC erzeugt, sondern ist vorgegeben:

$$A = \begin{pmatrix} 1-\epsilon & \epsilon & 0 & 0 & \dots & 0 \\ \epsilon & 1-2\epsilon & \epsilon & 0 & \dots & 0 \\ 0 & \epsilon & 1-2\epsilon & \epsilon & \dots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \dots & 0 & \epsilon & 1-2\epsilon & \epsilon \\ 0 & \dots & 0 & 0 & \epsilon & 1-\epsilon \end{pmatrix}$$

Was für einen Messprozess beschreibt die Matrix A? Im Folgenden sei $\epsilon = 0,23$. Implementieren sie eine Methode, mit der die Matrix A für beliebige Dimensionen $n \geq 3$ erzeugt werden kann.

b) Die wahre Verteilung f sei im Intervall $[0, 2]$ in 20 gleichgroßen Bins gegeben durch

$$f = [193, 485, 664, 763, 804, 805, 779, 736, 684, 626,$$
$$566, 508, 452, 400, 351, 308, 268, 233, 202, 173]$$

Nun wird eine Messung simuliert, indem zuerst die wahre Verteilung f mit A gefaltet wird: $g = A \cdot f$. Die „gemessenen“ Bin-Einträge der Verteilung g^{mess} erhält man durch Ziehen aus einer Poisson-Verteilung mit dem Erwartungswert des Bin-Eintrags aus der vorherigen Faltung: $\lambda_i = g_i$. Dadurch ergeben sich statistische Schwankungen in der Messung, die die Entfaltung erschweren.
Beispiel: Die Faltung ergibt einen Wert λ_i in Bin i der Verteilung g. Dann ergibt sich der gemessene Eintrag g_i^{mess} durch eine Zufallszahl, die aus einer Poisson-Verteilung mit Erwartungswert λ_i gezogen wird.

c) Transformieren Sie die Faltungsgleichung in die Diagonalbasis von $A = U \cdot D \cdot U^{-1}$. Wie lautet die Faltungsgleichung in der neuen Basis und welchen Vorteil bietet sie? Sortieren Sie die Eigenwerte absteigend nach ihrem Betrag und ändern Sie dementsprechend die Reihenfolge der zugehörigen Eigenvektoren in der Transformationsmatrix U.

d) Transformieren Sie die Verteilungen $f \to b$ und $g \to c$ mit der Matrix U aus **c)** in die neue Basis. Berechnen Sie ebenfalls die Kovarianzmatrix der Verteilung b mittels der „BVB“-Formel. Normieren Sie die entfalteten Koeffizienten $b = D^{-1}U^{-1}g^{\text{mess}}$ auf ihre Standardabweichungen und stellen Sie die Einträge b_j in einem Plot gegen ihren Index dar. Was lässt sich über Koeffizienten b_j sagen, die in dem Plot unterhalb der 1 liegen?
Info: Falls Sie die Verteilung g nicht erzeugen konnten, benutzen Sie hier die folgenden Werte:

$$g = [256, 457, 630, 787, 774, 816, 771, 749, 665, 647,$$
$$543, 510, 440, 410, 348, 296, 279, 243, 205, 174]$$

e) Regularisieren Sie, indem Sie nicht signifikante Koeffizienten b_j ab einen Cutoff-Index auf 0 setzen. Entfalten sie dann je einmal mit Regularisierung und ohne. Stellen Sie beide Lösungen mit statistischen Fehlern in einem Plot zusammen mit der gegebenen wahren Verteilung dar. Was ist der Unterschied der Lösung mit und ohne Regularisierung?

Aufgabe 38: *Entfaltung: Implementierung* **7 P.**

Zur Lösung dieser Aufgabe benötigen Sie die im moodle befindliche Datei `unfolding_data.hdf5`.

a) Lesen Sie den Key `train` aus der Datei `unfolding_data.hdf5` aus dem moodle ein und stellen Sie alle vorhandenen Attribute zusammen mit `energy_true` in jeweils einem zwei-dimensionalen Histogram dar. Wählen Sie das ihnen am geeignetsten erscheinende Attribut als Entfaltungs-Observable aus.

b) Berechnen Sie die Migrations-Matrix A. Nutzen Sie für die Observable 24 bins von 120 bis 500, und für die Zielvariable 16 bins von 15 bis 200. Stellen Sie die Migrationsmatrix grafisch dar.

c) Nehmen Sie eine Poisson-Statistik für die Werte in $\mathbf{g}$ an. Stellen Sie eine Likelihood-Funktion für die Verteilung der Zielvariable $\mathbf{f}$ auf.

d) Fügen Sie nun den Tikhonov-Regularisierungs-Term $\frac{\tau}{2}||\Gamma \cdot \mathbf{f}||^2$ hinzu. Berechnen Sie analytisch den Gradienten $(\nabla(-\log\mathcal{L}))(\mathbf{f})_i = \frac{\partial(-\log\mathcal{L})}{\partial f_i}$ und die Hesse-Matrix $\mathrm{H}[-\log\mathcal{L}](\mathbf{f})_{ij} = \frac{\partial^2(-\log\mathcal{L})}{\partial f_i \partial f_j}$

e) Implementieren Sie zur Minimierung der Likelihood das Newton-Verfahren gemäß der Rechenvorschrift

$$\mathbf{f}^{(k+1)} = \mathbf{f}^{(k)} - \left[\mathrm{H}[-\log\mathcal{L}](\mathbf{f}) + \epsilon\mathbf{1}\right]^{-1} \cdot (\nabla(-\log\mathcal{L}))(\mathbf{f}) \tag{1}$$

und nutzen Sie $\epsilon = 10^{-6}$. Brechen Sie das Verfahren ab, sobald $||\mathbf{f}^{(k+1)} - \mathbf{f}^{(k)}|| < 10^{-10}$ ist.

f) Laden Sie nun den Key `test` aus der Datei `unfolding_data.hdf5` aus dem moodle und berechnen Sie mit dem Binning aus b) die Vektoren $\mathbf{g}$ und $\mathbf{f}$. Entfalten Sie die Verteilung $\mathbf{g}$ und vergleichen Sie diese mit der wahren Verteilung $\mathbf{f}$ mit den Regularisierungsstärken $\tau = \{10^{-6}, 10^{-3}, 10^{-1}\}$. Welche Regularisierung erscheint angemessen?