Redukce dimenzionality

(a) *(2 b)* Definujte probléme redukce dimenzionalityy formálně (vstup, výstup, předpoklady, kritéria řešení).

Vstup:

$\mathbf{X} = \{\mathbf{x_i}\}_{i=1}^m \subset \mathcal{X}$ of dimension D (typically $\mathbb{R}^D$)

Výstup:

- a transformed space $\mathcal{T}$ of dimension L,
- dimensionality reduction mapping $\mathbf{F} : \mathcal{X} \to \mathcal{T}$,
- reconstruction mapping $\mathbf{f} : \mathcal{T} \to \mathcal{M} \subset \mathcal{X}$,

Předpoklady

$\mathbf{X}$ at least approximately lies on a manifold with $d < D$,

Manifold - je topologický prostor, který lokálně připomíná Euklidovský prostor

Kritéria řešení:

- $L < D$, L is as small as possible, at best $L = d$ (the intrinsic dimension),
- the manifold approximately contains all the sample points

$$\{\mathbf{x_i}\}_{i=1}^m \underset{\sim}{\subset} \mathcal{M} \overset{def}{=} f(\mathcal{T}),$$

- or alternatively, the reconstruction error of the sample is small

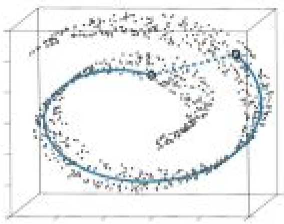
$$E_d(\mathbf{X}) \overset{def}{=} \sum_{i=1}^m d(\mathbf{x_i}, \mathbf{f}(\mathbf{F}(\mathbf{x_i}))).$$

(b) *(2 b)* Definujte a vysvětlete pojem multidimenzionální škálování.

Stejně jako PCA metoda se snaží redukovat dimenzionalitu. Hlavní rozdíl je, ale že místo aby se soustředila na korelaci mezi vzorky, tak se soustředí na vzdálenost mezi vzorky. Hlavní myšlenkou je, že body v originálním prostoru, které byly blízko by se měly i v prostoru s nižší dimenzí namapovat k sobě.

(c) *(2 b)* Definujte pojem geodetická vzdálenost (geodesic distance). Popište možnosti jejího využití při redukci dimenze.

Nejkratší délka cesty spojující dva body, která se nachází na manifoldu. Může nám vyřešit problémy, které se vyskytují s euklidovskou vzdáleností. Využívá se například v ISOMAP. Viz obrázek:



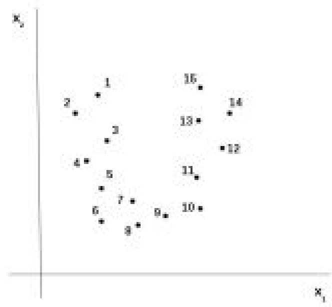
(d) *(2 b)* Napište pseudokód metody založené na multidimenzionálním škálování a geodesické vzdálenosti. Pojmenujte tuto metodu.

ISOMAP

1. Pro každý vzorek (bod) nalezneme nejbližší sousedy
   a. KNN nebo ve fixní vzdálenosti

2. Sestrojíme graf soused
    a. Každý bod spojím s jeho sousedy
    b. Délka hrany se rovná geodetické vzdálenosti mezi body
3. Nalezneme nejkratší cestu mezi všemi dvojicemi bodů
4. Sestrojíme pomocí MDS mapování nižší dimenze

(c) *(2 b)* Na obrázcích níže naznačte funkci metody popsané výše (tj. graficky naznačte způsob mapování bodů z prostoru vyšší dimenze do prostoru dimenze nižší). Může výstup vypadat i jinak než jste zakreslili?



Řešení může vypadat i jinak.

Multivariátní regrese. *(10 b)* Sestavujete multivariátní lineární model. Závisle proměnných je velký počet, jejich relevance je odlišná, některé z těchto proměnných jsou zcela irelevantní.

(a) *(2 b)* Pojmenujte 2 základní metody, pomocí kterých lze dosáhnout smrštění (shrinkage) výsledného modelu a zapište kriteriální funkce, které tyto dvě metody minimalizují.

LASSO

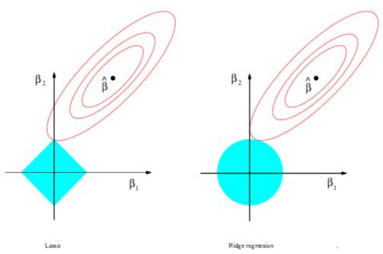
$$\sum_{i=1}^{m}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\hat{\beta}_j| = RSS + \lambda\sum_{j=1}^{p}|\hat{\beta}_j|$$

Ridge regression

$$\sum_{i=1}^{m}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^2 = RSS + \lambda\sum_{j=1}^{p}\hat{\beta}_j^2$$

(b) *(2 b)* Vysvětlete, v čem se výstup výše uvedených metod bude lišit. Zdůvodněte.

Ridge regression výsledek bude obsahovat všechny prediktor p, kvůli tomu se nezbavíme irelevantních proměnných. LASSO narozdíl od ridge regression bude mít ve výsledku některé koeficienty prediktorů nula a tím se jich zbaví. Důvodem je, že LASSO používá L1 normu, takže nutí některé koeficienty být nula zatímco ridge používá L2 normu. Viz obrázek:



(c) *(1 b)* Vyžaduje některá z výše uvedených metod předzpracování dat? Pokud ano, jaké a proč?

musíš si ty data standardizovat

posunout do středu, a normovat buď rozptylem a nebo normou protože kdybys měl jednu proměnnou v kilometrech a jednu v nanometrech, tak budou mít úplně jiný měřítko, ten koeficient beta u nanometrů bude mnohem větší než u kilometrů aby to vyvážil, a když to budeš regularizovat, tak bys jim oběma stáhnul hodnotu stejnym způsobem

(e) *(2 b)* Vysvětlete, proč mohou smrštěné modely dosáhnout nižší testovací chyby než referenční plný model vytvořený metodou nejmenších čtverců. Vysvětlení podpořte kompromisem mezi zaujetím (bias) a rozptylem (variance) obou typů modelů (plný vs smrštěný). Oba pojmy potřebné k vysvětlení definujte.

(f) *(1 b)* V čem jsou nevýhody smršťování oproti klasické aplikaci nejmenších čtverců?

nevýhody shrinkage metod jsou nutnost ladit parametr (lambda), a nejasná interpretace koeficientů, případně testy hypotéz nad nima

Robustní statistika

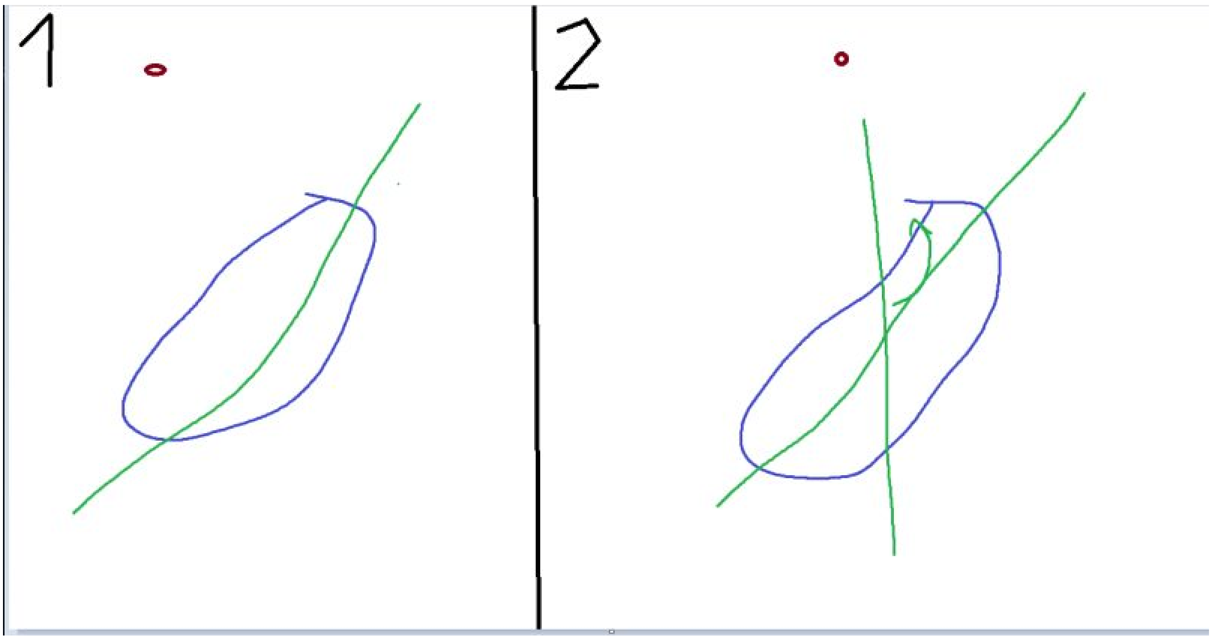
(a) *(2 b)* Co to je robustní regrese a za jakých podmínek je její využití vhodné?

neoptimalizuješ čtverce, ale nějakou robustnější ztrátu - sumu absolutních odchylek, nebo sumu Huber losses

(b) *(1 b)* Myšlenku robustní regrese demonstrujte graficky (postačí příklad jedné závisle a jedné nezávisle proměnné, bodový graf a srovnání výstupu robustní a klasické regrese).

nakreslíš dvakrát ten samej dataset, kde bude obláček pár hezkejch bodů a jeden zmrd puntík někde v piči

a ukážeš že přímka fitnutá obyčejnou regresí totálně uhne směrem k puntíku v piči (ale ne úplně na něj, jenom tim směrem), zatímco robustní přímka se ani nehne, případně jen velmi lehce



.

(c) *(2 b)* Jak lze regresní úlohu přeformulovat, aby šlo o robustní regresi? Popište alespoň 2 možnosti formulace.

neoptimalizuješ čtverce, ale nějakou robustnější ztrátu - sumu absolutních odchylek, nebo sumu Huber losses

(d) *(4 b)* Máte k dispozici dva párové vzorky:

$s_1 = \{293, 311, 331, 295, 337, 328, 291, 306, 323, 316\}$,

$s_2 = \{298, 322, 321, 321, 343, 331, 289, 316, 329, 322\}$.

Statisticky srovnejte oba vzorky na základě vhodné míry polohy (estimation of location, central tendency). Využijte jednu parametrickou a jednu neparametrickou, tj. robustní metodu. Pro obě metody formulujte jasný závěr.

Můžete využít těchto formulí: $T = \frac{\bar{d} - D_0}{s_d/\sqrt{n}}$, $W = \sum_{i=1}^{n}(sgn(x_i - y_i)R_i)$.

Příslušné tabulkové hodnoty: $t_{0.95,9} = 1.883$, $t_{0.975,9} = 2.262$, $t_{0.99,9} = 2.281$, $t_{0.995,9} = 3.250$;

$w_{0.95,10} = 40$, $w_{0.99,10} = 51$.

mean=313.1, 319.2
median=311-316, 321-322
atd…

t = 2.0842, tedy zamitame na hladine 0.95, ale dal ne
W = 38 a tak nezamitame H_0

.

(e) *(1 b)* Srovnejte výhody a nevýhody obou přístupů pro danou dvojici vzorků.

Power analysis

(a) *(2 b)* Vysvětlete pojem síla statistického testu.

Pravděpodobnost, že zamítnu H0, když Ha. Síla testu závisí na tom jak často se vyskytuje error Typu 2 (Beta). Beta nam říká jak moc je test nerozhodný. Malá Beta = test je nerozhodný, ale když už se rozhodne tak to je většinou pravda

(b) *(3 b)* Na čem síla testu závisí a jak (uveďte tři faktory, pro každý faktor popište typ vztahu)?

počet samplů = s počtem roste síla
alpha = závisí nepřímo
d = míra porušení H0, jistota zamítnutí -> tím větší tím lepší

(c) *(5 b)* Kolik účastníků testu je třeba pozvat na testování, pokud chcete mít 90% šanci vidět problémy, které postihují 30% všech uživatelů? Napište a popište rovnici pro výpočet velikosti vzorku pro objevování problémů. Proveďte výpočet.

n = log(1-x)/log(1-y) = log(1-0.9)/log(1-0.3) = 6.46, tudiz 7 lidi