Projekt 7: Övervakad maskininlärning: mönsterklassificering

15 mars 2017

Klassificering av mönster är en vital del av maskininlärning. I övervakad inlärning har man data från några olika sorters objekt (ofta bara två typer eller en handfull). Vi vill lära en dator att bestämma vilken typ en datavektor kommer från. För att lära datorn har vi tillgång till data där vi vet den korrekta klassificeringen. (Det faktum att vi har denna kunskap är orsaken till att vi kallar detta för övervakad inlärning.) Utöver kunskap om den korrekta klassificieringen, ger vi också datorn en modell för hur data skapas. Utifrån dessa kan datorn skapa en algoritm för klassificeringen. Syftet med att skapa en sådan algoritm är förstås att erhålla en algoritm som kan klassificera framtida data av samma typ.

En mycket vanlig modell är en Bayesiansk modell. Antag att vi har K typer av objekt, $1, 2, \ldots, K$. Låt A_k, $k = 1, \ldots, K$ vara händelsen att en ny observation kommer från typ k. För varje k har sedan data, X, för den givna observationen (om X är kontinuerlig) en betingad täthet $f(x|A_k)$.

Det vi observerar är $X = x$ och inte vilken typ observationen kommer från. Vi vill därför finna $\mathbb{P}(A_i|X = x)$. Enligt Bayes formel är

$$\mathbb{P}(A_k|X = x) \propto f(x|A_k)\mathbb{P}(A_k).$$

Naturligt är sedan att klassificera observationen som hörande till den typ k för vilken $\mathbb{P}(A_k|X = x)$ är störst. Om faktorerna på högersidan vore kända, skulle vi redan här ha en algoritm för klassificering. Men de tre sannolikheterna $\mathbb{P}(A_k)$ är okända, liksom eventuella parametrar hörande till de betingade tätheterna. Dessa måste skattas från data. När parametrarna är skattade har vi med detta skapat en algoritm för klassificering.

I det här projektet ska vi titta på ett konkret fall. Vi har tillgång till det s.k. Fisher's Iris Data Set, som ni laddar ner från kurshemsidan, med mätningar av några olika egenskaper hos irisblommor. Man har för varje observerad iris mätt längd och bredd på kronblad och foderblad, dvs en vektor i $\mathbb{R}^4$, och bestämt om aktuell blomma hör till en av tre arter: 1) setosa, 2) versicolor, 3) virginica. I modellen ska vi anta att en observation från en given art k kommer från en fyrdimensionell normalfördelning, med en väntevärdesvektor μ_k och en kovariansmatris Σ_k, dvs

$$f(x|A_k) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma_k}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right).$$

(Läs på om den multivariata normalfördelningen.)

Dela nu slumpvis upp data i två lika stora mängder D_1 och D_2, träningsdata respektive valideringsdata. Använd sedan data i D_1 för att skatta parametrarna och använd den erhållna klassificeringsalgoritmen på data i D_2. Eftersom ni vet den rätta klassificeringen i D_2 kan ni nu se hur stor andel av dessa data som klassificerades korrekt. Hur gick det?

Ett potentiellt problem ligger i att det är väldigt många parametrar att skatta med ganska lite data. Varje kovariansmatris har 10 parametrar och sannolikheterna för respektive art ger två parametrar till.

Det blir alltså 32 parametrar för 75 datapunkter och vi riskerar en överanpassning (vad betyder detta?) Pröva nu att anta att de tre kovariansmatriserna är lika: $\Sigma_k = \Sigma$. Skatta då Σ med medelvärdet av de tre individuella kovariansskattningarna. Då krymper antalet skattade parametrar till 12. Hur stor andel av blommorna i valideringsdata skattas nu rätt?

Slutligen några ord om datafilen. Den består dels av 150×4-matrisen `features` där varje rad svarar mot de fyra koordinaterna i respektive blommas data, dels av vektorn `class` där element k anger korrekt art för rad k i `features`.