Bayes-päättely
Työterveyslaitos
8.-9.2.2018

Ville Hyvönen

5. Mallinvalinta

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

1. Mallin yleistyminen
2. Testiaineisto
3. Ristiinvalidointi

Mallin yleistyminen

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Tilastollinen malli sovitetaan havaittuun aineistoon $\mathbf{y} = (y_1, \ldots, y_n)$.

Mallin yleistyminen

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Tilastollinen malli sovitetaan havaittuun aineistoon $\mathbf{y} = (y_1, \ldots, y_n)$.
- Lopullisena tavoitteena ei kuitenkaan ole saada mallia sopimaan mahdollisimman hyvin havaittuun aineistoon, vaan saada malli **yleistymään** mahdollisimman hyvin uusiin samaa jakaumaa noudattaviin aineistoihin

Mallin yleistyminen

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Tilastollinen malli sovitetaan havaittuun aineistoon $\mathbf{y} = (y_1, \ldots, y_n)$.
- Lopullisena tavoitteena ei kuitenkaan ole saada mallia sopimaan mahdollisimman hyvin havaittuun aineistoon, vaan saada malli **yleistymään** mahdollisimman hyvin uusiin samaa jakaumaa noudattaviin aineistoihin
- Tarkoituksena on siis kuvata havaitun aineiston generoinutta mekanismia, ei havaittuun aineistoon liittyvää satunnaisvaihtelua.
  - Signal vs. noise.

Mallin yleistyminen

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Formaalimmin tavoitteena ei siis ole löytää parametrin arvoa $\hat{\boldsymbol{\theta}}$ (oletetaan hetkeksi, että tiivistetään posteriorijakauma piste-estimaattiin $\hat{\boldsymbol{\theta}}$) joka maksoimisi aineiston todennäköisyyden $p(\mathbf{y}|\hat{\boldsymbol{\theta}})$

Mallin yleistyminen

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Formaalimmin tavoitteena ei siis ole löytää parametrin arvoa $\hat{\boldsymbol{\theta}}$ (oletetaan hetkeksi, että tiivistetään posteriorijakauma piste-estimaattiin $\hat{\boldsymbol{\theta}}$) joka maksoimisi aineiston todennäköisyyden $p(\mathbf{y}|\hat{\boldsymbol{\theta}})$
- ...vaan löytää parametrin arvo $\hat{\boldsymbol{\theta}}$, joka maksimoi (odotusarvomielessä) uuden samasta jakaumasta generoidun aineiston $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$ todennäköisyyden $p(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}})$.

Testiaineisto

- Paras tapa estimoida mallin yleistymistä, jos käytettävissä on erityinen testiaineisto $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$.

Testiaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Paras tapa estimoida mallin yleistymistä, jos käytettävissä on erityinen testiaineisto $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$.
- Lasketaan tn $p(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}})$ tälle testiaineistolle.

Testiaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Paras tapa estimoida mallin yleistymistä, jos käytettävissä on erityinen testiaineisto $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$.
- Lasketaan tn $p(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}})$ tälle testiaineistolle.
- Voidaan jakaa havaittu aineisto harjoitusaineistoon, johon mallit sovitetaan, ja testiaineistoon, jota käytetään niiden vertailuun.

Testiaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Paras tapa estimoida mallin yleistymistä, jos käytettävissä on erityinen testiaineisto $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_m)$.
- Lasketaan tn $p(\tilde{\mathbf{y}}|\hat{\boldsymbol{\theta}})$ tälle testiaineistolle.
- Voidaan jakaa havaittu aineisto harjoitusaineistoon, johon mallit sovitetaan, ja testiaineistoon, jota käytetään niiden vertailuun.
  - esim. 4/5 tai 9/10 aineistosta harjoitusaineistoksi ja 1/5 tai 1/10 testiaineistoksi.

Ristiinvalidointi

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Testiaineiston perusteella tehdyssä vertailussa myös satunnaisvaihtelua, riippuu training / test set - jaosta.

Ristiinvalidointi

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Testiaineiston perusteella tehdyssä vertailussa myös satunnaisvaihtelua, riippuu training / test set - jaosta.
- Ristiinvalidointi (cross-validation): Toistetaan testi-harjoitu-jako k kertaa (k-fold cross-validation).

Ristiinvalidointi

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Testiaineiston perusteella tehdyssä vertailussa myös satunnaisvaihtelua, riippuu training / test set - jaosta.
- Ristiinvalidointi (cross-validation): Toistetaan testi-harjoitu-jako k kertaa (k-fold cross-validation).
- Esim 5-kertaisessa ristiinvalidoinnissa jaetaan aineiston 5:een osaan, ja sovitetaan malli 5 kertaa aina 4/5 osaan aineistosta ja mitataan sen virhettä lopussa 1/5:ssa.

Ristiinvalidointi

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Testiaineiston perusteella tehdyssä vertailussa myös satunnaisvaihtelua, riippuu training / test set - jaosta.
- Ristiinvalidointi (cross-validation): Toistetaan testi-harjoitu-jako k kertaa (k-fold cross-validation).
- Esim 5-kertaisessa ristiinvalidoinnissa jaetaan aineiston 5:een osaan, ja sovitetaan malli 5 kertaa aina 4/5 osaan aineistosta ja mitataan sen virhettä lopussa 1/5:ssa.
- Lopuksi otetaan keskiarvo jokaisen viiden kerran virheestä.

Leave-one-out cross-validation

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Voidaan jopa jakaa aineisto n:ään osaan, jolloin sovitetaan malli n kertaa ja testataan sen virhettä yhdelle pois-jätetylle pisteelle kerrallaan.
- Tällöin puhutaan **leave-one-out cross-validaatiosta** (LOOCV).

Leave-one-out cross-validation

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Voidaan jopa jakaa aineisto n:ään osaan, jolloin sovitetaan malli n kertaa ja testataan sen virhettä yhdelle pois-jätetylle pisteelle kerrallaan.
- Tällöin puhutaan **leave-one-out cross-validaatiosta** (LOOCV).
- Laskennallisesti erittäin raskasta, malli sovitetaan m kertaa!

Leave-one-out cross-validation

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Voidaan jopa jakaa aineisto n:ään osaan, jolloin sovitetaan malli n kertaa ja testataan sen virhettä yhdelle pois-jätetylle pisteelle kerrallaan.
- Tällöin puhutaan **leave-one-out cross-validaatiosta** (LOOCV).
- Laskennallisesti erittäin raskasta, malli sovitetaan m kertaa!
- Yleensä 5- tai 10-kertainen ristiinvalidointi hyvä valinta.

Validaatioaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Jos testataan paljon malleja ristiinvalidoinnilla, ja valitaan niistä paras, saadaan helposti optimistinen arvio mallin yleistysvirheestä.

Validaatioaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Jos testataan paljon malleja ristiinvalidoinnilla, ja valitaan niistä paras, saadaan helposti optimistinen arvio mallin yleistysvirheestä.
- Ratkaisu: jaetaan aineisto validaatio- ja harjoitusaineistoon, esim. suhteessa 1/10 tai 1/5.

Validaatioaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Jos testataan paljon malleja ristiinvalidoinnilla, ja valitaan niistä paras, saadaan helposti optimistinen arvio mallin yleistysvirheestä.
- Ratkaisu: jaetaan aineisto validaatio- ja harjoitusaineistoon, esim. suhteessa 1/10 tai 1/5.
  - Valitaan ensin paras malli ristiinvalidoinnin perusteella.

Validaatioaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Jos testataan paljon malleja ristiinvalidoinnilla, ja valitaan niistä paras, saadaan helposti optimistinen arvio mallin yleistysvirheestä.
- Ratkaisu: jaetaan aineisto validaatio- ja harjoitusaineistoon, esim. suhteessa 1/10 tai 1/5.
  - Valitaan ensin paras malli ristiinvalidoinnin perusteella.
  - Sen jälkeen testataan tämän mallin yleistymisvirhettä validointiaineistolla, jota ei ole ollenkaan käytetty mallin opettamiseen.

Validaatioaineisto

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

- Jos testataan paljon malleja ristiinvalidoinnilla, ja valitaan niistä paras, saadaan helposti optimistinen arvio mallin yleistysvirheestä.
- Ratkaisu: jaetaan aineisto validaatio- ja harjoitusaineistoon, esim. suhteessa 1/10 tai 1/5.
  - Valitaan ensin paras malli ristiinvalidoinnin perusteella.
  - Sen jälkeen testataan tämän mallin yleistymisvirhettä validointiaineistolla, jota ei ole ollenkaan käytetty mallin opettamiseen.
  - Saadaan realistisempi arvio mallin yleistymisvirheestä.