Lecture 19

Slide 1

AIC
(Akaike's information criterion)

From relative entropy to maximum likelihood

- It has been shown that relative entropy (Kullback-Leibler divergence)

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

  can be used in model selection, if the 'true' model is known.

Slide 2

- If p is true probability source, we can choose q to be close to it in K-L sense.
- Usually the 'true' model is unknown. Is there a way to use K-L divergence in model selection?
- Next we show that there is a way, and it leads to an interesting connection with the maximum likelihood principle.

From relative entropy to maximum likelihood

- Relative entropy gives a 'distance' between two distributions

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) \tag{1}$$

Slide 3

- In Eq. 1 $p(x)$ is the true distribution of phenomenon under investigation. In practice, $p(x)$ is unknown but we may have observations about $p(x)$.
- $q(x)$ is the distribution which is compared to the true distribution. This could be a model to estimate a distribution of the phenomenon.

From relative entropy to maximum likelihood

- In equation $D(p||q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x)$ only the latter term on the right side of the equation depends on the model $q(x)$.

Slide 4

- The first term is a constant relative to the model selection problem $\Rightarrow$ we need to concentrate only on the second term.
- Notice: the second term is expected logarithmic loss.
- Next we'll see that the second term is the expected value of log likelihood of q.

From relative entropy to maximum likelihood

Slide 5

- Let's have a closer look of term $\sum_x p(x) \log q(x)$.
- Assume that we have N observations $(x_1, \ldots, x_N)$, where each observation has a value given by one of the m possible realizations $\omega_1, \ldots, \omega_m$.
- There is a true generating pdf $p(\omega_i) = \mathbb{P}(X = \omega_i),\ i = 1, \ldots, m$. The number of observed realizations for ω_i is described by variable n_i, which sum up to N; $(n_1 + \ldots + n_m = N)$.

From relative entropy to maximum likelihood

Slide 6

- Now we can rewrite the term $\sum_{i=1}^{m} p(\omega_i) \log q(\omega_i)$. Compare it to the definition of expected value

$$\mathbb{E}[Y] = \sum_y y p(y)$$

- $\log q(\omega_i) = \log q(X = \omega_i)$ is a realization for a random variable $\log q(X)$ and $p(x) = \mathbb{P}(X = \omega_i)$ is the probability of the realization.
- Now we can define the term $\sum_x p(x) \log q(x)$ as expected value of log likelihood.

From relative entropy to maximum likelihood

- Expected value of log likelihood?
- By the strong law of large numbers the average

$$\frac{1}{N}\sum_{l=1}^{N}\log q(x_l) \tag{2}$$

Slide 7

converges to expected value of log likelihood as $N \to \infty$.

- Because random variable $\log q(X)$ will have realization $\log q(X = \omega_i)$ n_i times, the Eq. 2 can be rewritten as

$$\frac{1}{N}\sum_{i=1}^{m} n_i \log q(\omega_i),$$

where $l(q) = \sum_{i=1}^{m} n_i \log q(\omega_i)$ is the log likelihood for model q.

From relative entropy to maximum likelihood

- Summa Summarum: Relative entropy can expressed with two terms. The first term is constant relative to a model selection problem and the second is the expected value of log likelihood of the model.

Slide 8

- Small K-L distance indicates a good model.
- Therefore: the higher the log likelihood, the better the model.
- The same holds for continuous probability densities. Treatment is similar as in the discrete case.

From relative entropy to maximum likelihood

Slide 9

- Example: Persons A and B predict again the performance of a football team. Person A predicts that the team wins 70% of its games and person B predicts 50%. After a couple of seasons we have statistics that the team won 65 games out of 100. Applying the log likelihood principle, who did best, A or B?
- log likelihood is now

$$l(q) = 65 \log q_1 + 35 \log q_2$$

- for A $q = [0.7\ 0.3]$ and for B $q = [0.5\ 0.5]$.
- Corresponding log likelihoods are -65.32 for A and -69.31 for B. So A got it better this time.

From relative entropy to maximum likelihood

Slide 10

- Example 2: We have ten observations $x =$ [-1.10, -0.40, -0.20, -0.02, 0.02, 0.71, 1.35, 1.46, 1.74, 3.89]. Which model class f_1 or f_2 is better for the given data:

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ f_2(x) &= \frac{1}{\pi(x^2+1)} \end{aligned}$$

- log likelihood for f_1: $-5 \ln 2\pi - 0.5 \sum_i x_i^2 = -21.2$
- log likelihood for f_2: $-10 \ln \pi - \sum_i \ln(x_i^2 + 1) = -19.2$
- So for this data model class f_2 is better in log likelihood sense

Akaike's information criterion, AIC

Slide 11

- Could it be possible to compare model classes in such a way that we fit a ML model for each class, and find out which of the model gives the largest value of likelihood function?
- No! Because maximum likelihood favors overfitting.
- The more complex to model is (for example, more parameters), the better fit and the higher likelihood function value we obtain.
- Therefore, we need some other method for choosing the model class. AIC (Akaike's information criterion or, originally, 'A information criterion') is one of the earliest attempts.

Akaike's information criterion, AIC

Slide 12

- Akaike's information criterion is based on the idea that minimizing the relative entropy between the 'true' distribution p and the tentative model q yields the optimal model,

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x)$$

- However, since p is unknown, it is impossible to calculate the relative entropy, or more precisely, the expected log loss (expected log likelihood) for the model (latter term).

Akaike's information criterion, AIC

Slide 13

- We have already shown that we may estimate the expected log likelihood with the maximized log likelihood function.
- This estimate has one problem: maximized log likelihood function depends on the data (realization of the random variable X), so the estimate will be biased.
- We may improve the situation by considering the mean expected log likelihood (MELL) function, so that the effect of different realizations is compensated.

AIC: spoiler

Slide 14

- Akaike's work show that
  - maximized log likelihood is a biased estimate of MELL
  - bias is asymptotically equal to k, the number of estimable parameters in the model
- Therefore, the AIC criterion for parametric model class selection is defined as

$$\mathrm{AIC}(k) = -2l(\hat{\theta}) + 2k \tag{3}$$

  where l is the log likelihood for the model, $\hat{\theta}$ are the ML estimates for the parameters and k is the number of parameters in the model. The lower the AIC, the better the model class.
- How on Earth can we obtain such a simple-looking formula from our starting point?

Derivation of AIC

- This sketch of derivation is neither rigorous nor detailed; only the main ideas are introduced.
- First we need a number of definitions.

Slide 15

- We assume that the data x has been generated by an unknown parametric density $f(x|\theta_0)$, where θ_0 is the true parameter vector.
- We define model classes $\mathcal{M}_k = \{ f(x|\theta^k) | \theta^k \in \Theta(k) \}$.
- Each class is a collection of densities parametrized by k-dimensional parameter vector $\theta^k = (\theta_1, \ldots, \theta_k)$.

Derivation of AIC

- We aim to fit a parametric model in such a way that we obtain a good approximation to $f(x|\theta_0)$.
- The likelihood function $f(x|\theta^k)$ is maximized by the ML parameters $\hat{\theta}^k$ for each model class.

Slide 16

- ML parameters are calculated always from observed data, $\hat{\theta}^k = \hat{\theta}^k(x_1, \ldots, x_n)$
- The log likelihood function is written as $l(\theta^k)$, and the corresponding maximized log likelihood $l(\hat{\theta}^k)$.

Derivation of AIC

- The expected log likelihood for any model is defined as

$$\mathbb{E}_{\theta_0}\left[\log f(X|\theta^k)\right] = \int f(x|\theta_0) \log f(x|\theta^k) dx \tag{4}$$

where $\mathbb{E}_{\theta_0}$ is an expectation with respect to the true density $f(x|\theta_0)$.

Slide 17

- For the maximum likelihood model, the expected (maximized) log likelihood is then of course

$$\mathbb{E}_{\theta_0}\left[\log f(X|\hat{\theta}^k)\right] = \int f(x|\theta_0) \log f(x|\hat{\theta}^k) dx \tag{5}$$

where ML parameters $\hat{\theta}^k$ are calculated with some given and fixed set of observations.

Derivation of AIC

- Expected maximum log likelihood depends on the observations (a single realization of random variable X) used to calculate the maximum likelihood estimates.

Slide 18

- We may try to improve the quality of our evaluation by considering the *mean* expected maximum log likelihood.
- Akaike: when the *mean expected maximum log likelihood*

$$\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k(Y))\right] \tag{6}$$

gets large value, the better the model is.

Derivation of AIC

Slide 19

- Mean expected maximum log likelihood (MELL):

$$\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k(Y))\right]$$

- Although only Y represents the data (which is used to calculate ML estimates), it is convenient to conceptualize X and Y as i.i.d. samples from the same true distribution.
- MELL still depends on the true distribution, so we have to find a way to estimate it.
- Next we shall see that maximum likelihood function is a biased estimate for MELL, and that the bias can be obtained easily as the number of free parameters in the model.

Derivation of AIC

Slide 20

- How to estimate $\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}[\log f(X|\hat{\theta}^k(Y))]$?
- It can be done in two parts
  1. Estimate difference between $\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}[\log f(X|\hat{\theta}^k(Y))]$ and $\mathbb{E}_{X|\theta_0}[\log f(X|\theta_0)]$
  2. Estimate difference between $\mathbb{E}_{X|\theta_0}[\log f(X|\theta_0)]$ and $\log f(X|\hat{\theta}^k)$.
- Combine results from 1. and 2. to give the complete estimator.

Derivation of AIC

- Part 1. We can show that

$$\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k(Y))\right] = \mathbb{E}_{X|\theta_0}\left[\log f(X|\theta_0)\right] - \frac{k}{2} \qquad (7)$$

- Idea for proof:

Slide 21

  - Take the 2nd order Taylor approximation of $\mathbb{E}_{X|\theta_0}[\log f(X|\theta^k)]$ around the true parameters θ_0.
  - 0th order term is $\mathbb{E}_{X|\theta_0}[\log f(X|\theta_0)]$.
  - 1st order term vanishes, because expected log likelihood is maximized at the true parameters.
  - 2nd order term is given by $\frac{1}{2}\sqrt{n}(\theta^k - \theta_0)J(\theta_0)\sqrt{n}(\theta^k - \theta_0)^T$.

Derivation of AIC

- Idea for proof continues:

Slide 22

  - The quadratic term $\sqrt{n}(\theta^k - \theta_0)J(\theta_0)\sqrt{n}(\theta^k - \theta_0)^T$ converges to a centrally distributed χ^2 random variable with k degrees of freedom.
  - Take the expectation of the Taylor expansion to get the MELL
  - First remaining term is $\mathbb{E}_{X|\theta_0}[\log f(X|\theta_0)]$
  - Expectation of the χ^2 term is $\mathbb{E}_{\theta_0}\sqrt{n}(\theta^k - \theta_0)J(\theta_0)\sqrt{n}(\theta^k - \theta_0)^T = k$

Derivation of AIC

Slide 23

- All said, we get an asymptotic approximation

$$\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k(Y))\right] = \mathbb{E}_{X|\theta_0}\left[\log f(X|\theta_0)\right] - \frac{k}{2} \quad (8)$$

- For part 2 we get with a similar treatment

$$\mathbb{E}_{X|\theta_0}\left[\log f(X|\theta_0)\right] = \mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k)\right] - \frac{k}{2} \quad (9)$$

Derivation of AIC

Slide 24

- By combining (8) and (9), we get an approximation for MELL:

$$\mathbb{E}_{Y|\theta_0}\mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k(Y))\right] = \mathbb{E}_{X|\theta_0}\left[\log f(X|\hat{\theta}^k)\right] - k \quad (10)$$

- This means that maximum likelihood is a biased estimate for mean expected maximum likelihood, and that the bias is equal to the number of parameters in the model.
- In order to remove the bias we introduce an unbiased estimator for mean expected log likelihood

$$\log f(X|\hat{\theta}^k) - k \quad (11)$$

- This is how we arrive to Akaike's information criterion.

Akaike's Information Criterion

Slide 25

- AIC is usually presented as

$$\mathrm{AIC}(k) = -2l(\hat{\theta}) + 2k, \tag{12}$$

  for 'historical' reasons.

- When using this form for selecting the parametric model class, choose k for which the AIC(k) is lowest.

About AIC

Slide 26

- Some issues about AIC
  1. AIC is asymptotic; it requires conventional large-sample properties.
  2. The maximum number of parameters m_{max} should not exceed $2\sqrt{n}$, where n is the number of observation. This is because larger m_{max} weakens the bias correction.
  3. There are cases when AIC decreases monotonically, i.e., there is no solution. In most of these cases the culprit is poor selection of model class.
  4. If AIC score difference between two models is in magnitude of 1-2, the difference is significant.
  5. In some cases AIC has been shown to be inconsistent.

Example: AIC and linear regression

- Example: consider linear regression with m regressor variables.

$$\begin{aligned} y &= w_0 + \sum_{i=1}^{m} w_i x_i + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

Slide 27

- AIC: $-2l(\hat{\theta}) + 2k$
- Pdf for a single observation

$$p(y|\mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y - w_0 - \sum_{i=1}^{m} w_i x_i)^2}$$

- Likelihood for n i.i.d. observations:

$$L(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} p(y_i|\mathbf{w}, \sigma^2)$$

Example: AIC and linear regression

- Log likelihood:

Slide 28

$$l(\mathbf{w}, \sigma^2) = -\frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - w_0 - \sum_{j=1}^{m} w_j x_{ij} \right)^2$$

Example: AIC and linear regression

Slide 29

- AIC:

$$\mathrm{AIC}(k) = -2l(\hat{\theta}) + 2k$$

- The number of parameters in this model is $m + 2$ ($w_0, \ldots, w_m, \sigma^2$).
- Solve the maximum likelihood estimates and substitute to AIC criterion

$$\mathrm{AIC}(k) = n\ln(2\pi) + n + n\ln(\hat{\sigma}^2) + 2(m+2) \qquad (13)$$