08_gaussian_ch8

Randy

9/23/2021

$\mathcal{O}(.)$ problem

- A significant problem with Gaussian process prediction is that it typically scales as $\mathcal{O}(n^3)$.
- prohibitive problems for large dataset (e.g. $n > 10,000$):
  - storing the Gram matrix
  - solving the associated linear systems

Inversion Lemma

To invert the matrix $K + \sigma_n^2 I$ (or at least to solve a linear system $(K + \sigma_n^2 I)\boldsymbol{v} = \boldsymbol{y}$ for $\boldsymbol{v}$)

- K has rank q (so that it can be represented in the form $K = QQ^\top$;
- where Q is an $n \times q$ matrix)
- Matrix inversion can be speeded up using the matrix inversion lemma eq. (A.9)

$$(Z + UWV^\top)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^\top Z^{-1}U)^{-1}V^\top Z^{-1} \qquad (A.9)$$

- Result as $(QQ^\top + \sigma_n^2 I_n)^{-1} = \sigma_n^{-2} I_n - \sigma_n^{-2} Q(\sigma_n^2 I_q + Q^\top Q)^{-1} Q^\top$.

Notice that the inversion of an $n \times n$ matrix has now been transformed to the inversion of a $q \times q$ matrix.

Inversion Lemma for the kernel with N features

- The Gram matrix will have rank $min(n, N)$ so that exploitation of this structure will be beneficial if $n > N$.
- Even if the kernel is non-degenerate it may happen that it has a fast-decaying eigen-spectrum
- so that a reduced-rank approximation will be accurate

Even K is not of $rank < n$

- still consider reduced-rank approximations to K
- with the optimal reduced-rank approximation of K w.r.t. the **Frobenius norm** (see eq. (A.16))

$$\|A\|_F^2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|a_{ij}|^2 = tr(AA^\top) \qquad (A.16)$$

- $U_q\Lambda_q U_q^\top$ with Λ_q is the diagonal matrix with the first q eigenvalues of K and U_q is the matrix of the corresponding orthonormal eigenvectors
- Limit of computing the eigen-decomposition is an $\mathcal{O}(n^3)$ operation
- However, it does suggest that if we can more cheaply obtain an approximate eigen-decomposition (may give rise to a reduced rank approximation)

Setting up an active set

A subset I of the original n data points, called the **active set**.

- Setting I as size $m < n$ (I is for the included data point)
- Remaining $n - m$ data points form the set R (R is for the remaining points)
- WOLOG: the data points are ordered so that set I comes first
- K can be partitioned as

$$K = \begin{pmatrix} K_{mm} & K_{m(n-m)} \\ K_{(n-m)m} & K_{(n-m)(n-m)} \end{pmatrix} \qquad (8.1)$$

To approximate the eigenfunctions of a kernel using the Nystrom method.

- Compute the eigenvectors and eigenvalues of K_{mm} and denote them $\{\lambda_i^{(m)}\}_{i=1}^m$ and $\{\boldsymbol{u}_i^{(m)}\}_{i=1}^m$.
- Extended to all n points using **eq. (4.44)** $(\boldsymbol{k}(\boldsymbol{x}') = \{_C\ k(\boldsymbol{x}_i,\ \boldsymbol{x}')\}_{i=1}^n)$

$$\phi_i(\boldsymbol{x}') \simeq \frac{\sqrt{n}}{\lambda_i^{mat}} \boldsymbol{k}(\boldsymbol{x}')^\top \boldsymbol{u}_i \qquad (4.44)$$

- $\tilde{\lambda}_i^{(n)} \triangleq \frac{n}{m}\lambda_i^{(m)},\ i = 1, ..., m \qquad (8.2)$
- $\tilde{\boldsymbol{u}}_i^{(n)} \triangleq \sqrt{\frac{m}{n}} \frac{1}{\lambda_i^{(m)}} K_{nm} \boldsymbol{u}_i^{(m)},\ i = 1, ..., m \qquad (8.3)$
- with the scaling of $\tilde{\boldsymbol{u}}_i^{(n)}$ has been chosen so that $|\tilde{\boldsymbol{u}}_i^{(n)}| \simeq 1$

Nystrom approximation to K

In general we can choice the approximate eigenvalues/vectors to include in approximation of K

- Choosing the first p values: $\tilde{K} = \sum_{i=1}^{p} \tilde{\lambda}_i^{(n)} \tilde{\boldsymbol{u}}_i^{(n)} (\tilde{\boldsymbol{u}}_i^{(n)})^\top$
- Now set $p = m$ to obtain $\tilde{K} = K_{nm} K_{mm}^{-1} K_{mn}$ (8.4)
- Combining equations 8.2, 8.3, and 8.4

Computation of $\tilde{K}$ takes time $\mathcal{O}(m^2 n)$

- The eigen-decomposition of K_{mm} is $\mathcal{O}(m^3)$
- The computation of each $\tilde{\boldsymbol{u}}_i^{(n)}$ is $\mathcal{O}(mn)$
- Up to ($10^6 \times 10^6$ in size by **Fowlkes et al. [2001]**).

The Nystrom approximation has been applied above to approximate the elements of K.

However, using the approximation for the ith eigen-function

$\tilde{\phi}_i(\boldsymbol{x}) = (\sqrt{m}/\lambda_i^{(m)})\boldsymbol{k}_m(\boldsymbol{x})^\top \boldsymbol{u}_i^{(m)}$ with $(\boldsymbol{k}(\boldsymbol{x}') = \{_C\, k(\boldsymbol{x}_i,\, \boldsymbol{x}')\}_{i=1}^n)$

- $\lambda_i' \simeq \lambda_i^{(m)}/m$ it is easy to see that in general we obtain an approximation for the kernel $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^N \lambda_i \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x}')$ as

$$\tilde{k}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^m \frac{\lambda_i^{(m)}}{m} \tilde{\phi}_i(\boldsymbol{x})\tilde{\phi}_i(\boldsymbol{x}') \qquad (8.5)$$

$$= \sum_{i=1}^m \frac{\lambda_i^{(m)}}{m} \frac{m}{(\lambda_i^{(m)})^2} \boldsymbol{k}_m(\boldsymbol{x})^\top \boldsymbol{u}_i^{(m)} (\boldsymbol{u}_i^{(m)})^\top \boldsymbol{k}_m(\boldsymbol{x}') \qquad (8.6)$$

$$= \boldsymbol{k}_m(\boldsymbol{x})^\top K_{mm}^{-1} \boldsymbol{k}_m(\boldsymbol{x}') \qquad (8.7)$$

- By multiplying out eq. (8.4) using $K_{mn} = [K_{mm} K_{m(n-m)}]$ it is easy to show that $K_{mm} = \tilde{K}_{mm}$, $K_{m(n-m)} = \tilde{K}_{m(n-m)}$, $K_{(n-m)m} = \tilde{K}_{(n-m)m}$, but that $\tilde{K}_{(n-m)(n-m)} = K_{(n-m)m} K_{mm}^{-1} K_{m(n-m)}$.
- The difference $K_{(n-m)(n-m)} - \tilde{K}_{(n-m)(n-m)}$ is in fact the **Schur complement** of $K_m m$ [Golub and Van Loan, 1989, p. 103].
- $K_{(n-m)(n-m)} - \tilde{K}_{(n-m)(n-m)}$ is positive semi-definite;
- If a vector $\boldsymbol{f}$ is partitioned as $\boldsymbol{f}^\top = (\boldsymbol{f}_m^\top, \boldsymbol{f}_{n-m}^\top)$ and $\boldsymbol{f}$ has a Gaussian distribution with zero mean and covariance K then $\boldsymbol{f}_{n-m} | \boldsymbol{f}_m$ has the **Schur complement** as its covariance matrix, see eq. (A.6).

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix} \right) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C}^\top & \tilde{B} \end{bmatrix}^{-1} \right) \qquad (A.5)$$

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, A)$$

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}(\mu_x + CB^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y),\ A - CB^{-1}C^\top \qquad (A.6a)$$

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}(\mu_x - \tilde{A}^{-1}\tilde{C}(\boldsymbol{y} - \boldsymbol{\mu}_y),\ \tilde{A}^{-1}) \qquad (A.6b)$$

An alternative view

- **Nystrom approximation** was derived in the above fashion by **Williams and Seeger [2001]** for application to kernel machines
- The same approximation is due to **Smola and Sch¨olkopf [2000]**
- To approximate the kernel centered on point $\boldsymbol{x}_i$ as a linear combination of kernels from the active set

$$\boldsymbol{k}(\boldsymbol{x}_i, \boldsymbol{x}) \simeq \sum_{j \in I} c_{ij} k(\boldsymbol{x}_j, \boldsymbol{x}) \stackrel{\triangle}{=} \hat{k}(\boldsymbol{x}_i, \boldsymbol{x}) \qquad (8.8)$$

for some coefficients $\{c_{ij}\}$ that are to be determined so as to optimize the approximation

A reasonable criterion to minimize

$$E(C) = \sum_{i=1}^{n} \|k(\boldsymbol{x}_i, \boldsymbol{x}) - \hat{k}(\boldsymbol{x}_i, \boldsymbol{x})\|^2_{\mathcal{H}} \qquad (8.9)$$

$$= tr\ K - 2tr(CK_{mn}) + tr(CK_{mm}C^\top) \qquad (8.10)$$

- The coefficients are arranged into a $n \times m$ matrix C.
- Minimizing $E(C)$ w.r.t. C gives $C_{opt} = K_{nm}K_{mm}^{-1}$
- Thus we obtain the approximation $\hat{K} = K_{nm}K_{mm}^{-1}K_{mn}$ in agreement with eq. (8.4).
- $E(C_{opt}) = tr(K - \hat{K})$

Smola and Scholkopf [2000] suggest a **greedy algorithm** to choose points to include into the **active set** so as to minimize the error criterion.

- $\mathcal{O}(mn)$ operations to evaluate the change in E due to including one new datapoint
- it is infeasible to consider all members of set R for inclusion on each iteration
- instead **Smola and Scholkopf [2000]** suggest finding the best point to include from a randomly chosen subset of set R on each iteration.

Drineas and Mahoney [2005] used biased sampling with replacement

- choosing column i of K with probability $\propto k_{ii}^2$
- a pseudoinverse of the inner $m \times m$ matrix

To provide probabilistic bounds on the quality of the approximation

Frieze et al. [1998] had developed an approximation to the singular value decomposition (SVD) of a rectangular matrix

- using a weighted random subsampling of its rows and columns, and probabilistic error bounds.
- However, this is rather different from the Nystrom approximation

** Fine and Scheinberg [2002]** suggest an alternative low-rank approximation to K using the incomplete Cholesky factorization

- when computing the Cholesky decomposition of K pivots below a certain threshold are skipped.
- If the number of pivots greater than the threshold is k the incomplete Cholesky factorization takes time $\mathcal{O}(nk^2)$

8.2 Greedy Approximation

- an active set of training points of size m selected from the training set of size $n > m$
- assume that it is impossible to search for the optimal subset of size m due to combinatorics.
- The points in the active set could be selected randomly
- but if the points are selected greedily w.r.t. some criterion, the results are better.
- greedy approaches are also known as forward selection strategies.

Algorithm

input: m, desired size of active set
Initialization $I = \emptyset$, $R = \{1, \ldots, n\}$
for $j := 1 \ldots m$ **do**
  Create working set $J \subseteq R$
  Compute Δ_j for all $j \in J$
  $i = \text{argmax}_{j \in J} \Delta_j$
  Update model to include data from example i
  $I \leftarrow I \cup \{i\}$, $R \leftarrow R \backslash \{i\}$
end for
return: I

Algorithm 8.1: General framework for greedy subset selection. Δ_j is the criterion function evaluated on data point j.

This is achieved by evaluating some criterion Δ and selecting the data point that optimizes this criterion.

For some algorithms it can be too expensive to evaluate Δ on all points in R, so some working set $J \subset R$ can be chosen instead, usually at random from R.

Greedy selection is used with the **subset of regressors (SR)**, **subset of datapoints (SD)** and the **projected process (PP)** methods

Total six approximation schemes for GPR below:

- The subset of regressors (SR)
- The Nystrom method
- The subset of datapoints (SD)
- The projected process (PP) approximation
- The Bayesian committee machine (BCM)
- The iterative solution of linear systems

8.3.1 Subset of Regressors

Silverman [1985, sec. 6.1] showed:

- the mean GP predictor can be obtained from a finite-dimensional generalized linear regression model
- $f(\boldsymbol{x}_*) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_*, \boldsymbol{x}i)$ with a prior $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, K^{-1})$
- the mean prediction for linear regression model in feature space given by eq. (2.11),

$$\bar{f}(\boldsymbol{x}_*) = \sigma_n^{-2} \phi(\boldsymbol{x}_*)^\top A^{-1} \Phi y$$
$$A = \Sigma_p^{-1} + \sigma_n^{-2} \Phi \Phi^\top$$
$$\phi(\boldsymbol{x}_*) = \boldsymbol{k}(x_*)$$
$$\Phi = \Phi^\top = K$$
$$\Sigma_p^{-1} = K$$

$$\bar{f}(\boldsymbol{x}_*) = \sigma_n^{-2}\boldsymbol{k}^\top(\boldsymbol{x}_*)[\sigma_n^{-2}K(K+\sigma_n^2 I)]^{-1}K\boldsymbol{y} \quad (8.11)$$
$$= \boldsymbol{k}^\top(\boldsymbol{x}_*)(K+\sigma_n^2 I)^{-1}\boldsymbol{y} \quad (8.12)$$

- this result is in agreement with eq. (2.25)
- however, that the predictive (co)variance of this model is different from full GPR.

A simple approximation to this model is to consider only a subset of regressors, so that

$$f_{SR}(\boldsymbol{x}_*) = \sum_{i=1}^{m} \alpha_i k(\boldsymbol{x}_*, \boldsymbol{x}_i), \ \alpha_m \sim \mathcal{N}(\boldsymbol{0}, K_{mm}^{-1}) \quad (8.13)$$

by $f_*|\boldsymbol{x}_*, X, \boldsymbol{y} \sim \mathcal{N}(\frac{1}{\sigma_n^2}\phi(\boldsymbol{x}_*)^\top A^{-1}\Phi\boldsymbol{y}, \ \phi(\boldsymbol{x}_*)T\top A^{-1}\phi(\boldsymbol{x}_*))$ (2.11)

$$\bar{f}_{SR}(\boldsymbol{x}_*) = \boldsymbol{k}_m(\boldsymbol{x}_*)^\top (K_{mn}K_{nm} + \sigma_n^2 K_{mm})^{-1} K_{mn}\boldsymbol{y} \quad (8.14)$$

$$\mathbb{V}[f_{SR}(\boldsymbol{x}_*)] = \sigma_n^2 \boldsymbol{k}_m(\boldsymbol{x}_*)^\top (K_{mn}K_{nm} + \sigma_n^2 K_{mm})^{-1} \boldsymbol{k}_m(\boldsymbol{x}_*) \quad (8.15)$$

$$\bar{\alpha}_m = (K_{mn}K_{nm} + \sigma_n^2 K_{mm})^{-1} K_{mn}\boldsymbol{y} \quad (8.16)$$

"subset of regressors" (SR) was suggested to us by G. Wahba.

- The computations for equations 8.14 and 8.15 take time $\mathcal{O}(m^2 n)$ to carry out the necessary matrix computations.
- After this the prediction of the mean for a new test point takes time $\mathcal{O}(m)$, and the predictive variance takes $\mathcal{O}(m^2)$.

Under the subset of regressors model we have $f \sim \mathcal{N}(0, \tilde{K})$ where $\tilde{K}$ is defined as in eq. (8.4).

Thus the log marginal likelihood under this model is $\log p_{SR}(y|X) = -\frac{1}{2}\log|\tilde{K} + \sigma_n^2 I_n| - \frac{1}{2}\ pmby^\top(\tilde{K} + \sigma_n^2 I_n)^{-1}\boldsymbol{y} - \frac{n}{2}\log(2\pi) \quad (8.17)$

Notice that the covariance function defined by the SR model has the form $\tilde{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{k}(\boldsymbol{x})^\top K_{mm}^{-1}\boldsymbol{k}(\boldsymbol{x}')$, which is exactly the same as that from the Nystrom approximation for the covariance function eq. (8.7).

In fact if the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ in the predictive mean and variance equations 2.25 and 2.26 is replaced systematically with $\tilde{K}(\boldsymbol{x}, \boldsymbol{x}')$ we obtain equations 8.14 and 8.15, as shown in Appendix 8.6.

If the kernel function decays to zero for $|x| \to \infty$ for fixed $\boldsymbol{x}'$, then $\tilde{K}(\boldsymbol{x}, \boldsymbol{x})$ will be near zero when $\boldsymbol{x}$ is distant from points in the set I.

This will be the case even when the kernel is stationary so that $k(x, x)$ is independent of x.

Thus we might expect that using the approximate kernel will give poor predictions, especially underestimates of the predictive variance, when $\boldsymbol{x}$ is far from points in the set I.

An interesting idea suggested by **Rasmussen and Quinonero-Candela [2005]** to mitigate this problem

- to define the SR model with $m+1$ basis functions, where the extra basis function is centered on the test point x_*
- so that $y_{SR*}(\boldsymbol{x}_*) = \sum_{i=1}^{m} \alpha_i k(\boldsymbol{x}_*, \boldsymbol{x}_i) + \alpha_* k(\boldsymbol{x}_*, \boldsymbol{x}_*)$.
- This model can then be used to make predictions, and it can be implemented efficiently using the partitioned matrix inverse equations A.11 and A.12.

The effect of the extra basis function centered on $\boldsymbol{x}_*$ is to maintain predictive variance at the test point.

- One simple method is to choose subset I randomly from X
- Another is to run clustering on $\boldsymbol{x}_i{}_{i=1}^{n}$ to obtain centers.
- Alternatively, a number of greedy forward selection algorithms for I have been proposed:
  - **Luo and Wahba [1997]** choose the next kernel so as to minimize the residual sum of squares (RSS) $|y - K_{nm}\boldsymbol{\alpha}_m|^2$ after optimizing $\boldsymbol{\alpha}_m$
  - **Smola and Bartlett [2001]** choose as their criterion the quadratic form

$$\frac{1}{2\sigma_n^2}|\boldsymbol{y} - K_{nm}\bar{\boldsymbol{\alpha}}_m|^2 + \bar{\boldsymbol{\alpha}}_m^\top K_{mm}\bar{\boldsymbol{\alpha}}_m = \boldsymbol{y}^\top(\tilde{K} + \sigma_n^2 I_n)^{-1}\boldsymbol{y} \quad (8.18)$$

- Alternatively, **Quinonero-Candela [2004]** suggests using the approximate $\log p_{SR}(y|X)$ (see eq. (8.17)) as the selection criterion.
- the quadratic term from eq. (8.18) is one of the terms comprising $\log p_{SR}(y|X)$.
- For all these suggestions the complexity of evaluating the criterion on a new example is $\mathcal{O}(mn)$, by making use of partitioned matrix equations.
- Thus it is likely to be too expensive to consider all points in R on each iteration
- Note that the SR model is obtained by selecting some subset of the data points of size m in a random or greedy manner.
- The relevance vector machine (RVM) described in section 6.6 has a similar flavour
- it automatically selects (in a greedy fashion) which data points to use in its expansion.

8.3.2 The Nystrom Method for approximate GPR

Williams and Seeger [2001] suggested:

- approximating the GPR equations by replacing the matrix K by $\tilde{K}$ in the mean and variance prediction equations 2.25 and 2.26.
- in this proposal the covariance function k is not systematically replaced by $\tilde{k}$
- it is only occurrences of the matrix K that are replaced.
- As for the SR model the time complexity is $\mathcal{O}(m^2 n)$ to carry out the necessary matrix computations
- then $\mathcal{O}(n)$ for the predictive mean of a test point
- $\mathcal{O}(mn)$ for the predictive variance

Experimental evidence in **Williams et al. [2002]** suggests:

- for large m the SR and Nystrom methods have similar performance
- but for small m the Nystrom method can be quite poor
- Also the fact that k is not systematically replaced by $\tilde{K}$ means that the approximated predictive variance might be negative.
- For these reasons, we do not recommend the Nystrom method over the SR method.
- However, the Nystrom method can be effective when λ_{m+1}, the $(m+1)$th eigenvalue of K, is much smaller than σ_n.

8.3.3 Subset of Datapoints

- to keep the GP predictor, but only on a smaller subset of size m of the data.
- Although this is clearly wasteful of data, it can make sense if the predictions obtained with m points are sufficiently accurate for our needs.
- it can make sense to select which points are taken into the active set I, and typically this is achieved by greedy algorithms.
- However, one has to be wary of the amount of computation that is needed, if one considers each member of R at each iteration.

Lawrence et al. [2003] suggest:

- the next point active set point can maximize the differential entropy score $\Delta_j \triangleq H[p(f_j)] - H[p^{new}(f_j)]$
- where $H[p(f_j)]$ is the entropy of the Gaussian at site $j \in R$ (which is a function of the variance at site j as the posterior is Gaussian, see eq. (A.20))
- $H[p^{new}(f_j)]$ is the entropy at this site once the observation at site j has been included.
- Let the posterior variance of f_j before inclusion be v_j.
- As $p(f_j|\boldsymbol{y}_I, y_j) \propto p(f_j|\boldsymbol{y}_I)N(y_j|f_j, \sigma^2)$ we have $(v_j^{new})^{-1} = v_j^{-1} + \sigma^{-2}$.

- Using the fact that the entropy of a Gaussian with variance v is $\log(2\pi ev)/2$

$$\Delta_j = \frac{1}{2}\log\left(1 + \frac{v_j}{\sigma^2}\right) \qquad (8.19)$$

- Δ_j is a monotonic function of v_j so that it is maximized by choosing the site with the largest variance.
- **Lawrence et al. [2003]** call their method the informative IVM vector machine (IVM)

- Coded naively computing the variance at all sites in R cost $\mathcal{O}(m^3 + (n-m)m^2)$ as we need to evaluate eq. (2.26) at each site
- the matrix inversion of $K_{mm} + \sigma_n^2 I$ can be done once in $\mathcal{O}(m^3)$ then stored.
- However, as we are incrementally growing the matrices K_{mm} and $K_{m(n-m)}$ in fact the cost is $\mathcal{O}(mn)$ per inclusion
- leading to an overall complexity of $\mathcal{O}(m^2 n)$ when using a subset of size m.

For example, once a site has been chosen for inclusion the matrix $K_{mm} + \sigma_n^2 I$ is grown by including an extra row and column.

- The inverse of this expanded matrix can be found using eq. (A.12) although it would be better practice numerically to use a Cholesky decomposition approach as described in Lawrence et al. [2003].

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix} \qquad (A.11)$$

$$\begin{cases} \tilde{P} = P^{-1} + P^{-1}QMRP^{-1} \\ \tilde{Q} = -P^{-1}QM \\ \tilde{R} = -MRP^{-1} \\ \tilde{S} = M \end{cases} \qquad M = (S - RP^{-1}Q)^{-1} \qquad (A.12)$$

- The scheme evaluates Δ_j over all $j \in R$ at each step to choose the inclusion site.
- This makes sense when m is small, but as it gets larger it can make sense to select candidate inclusion sites from a subset of R.
- **Lawrence et al. [2003]** call this the **randomized greedy selection method** and give further ideas on how to choose the subset.
  - The differential entropy score Δ_j is not the only criterion that can be used for site selection.
  - For example the information gain criterion $KL(p^{new}(f_j)||p(f_j))$ can also be used.
- The use of greedy selection heuristics here is similar to the problem of active learning, see e.g. **MacKay [1992c]**.