Lecture topics: model selection criteria

- Structural risk minimization, example derivation
- Bayesian score, Bayesian Information Criterion (BIC)

Model selection criteria: structural risk minimization

One perspective to model selection is to find the model (set of discriminant functions) that has the best *guarantee of generalization*. To obtain such guarantees we have to relate the *empirical risk* $R_n(\hat{f}_i)$

$$R_n(\hat{f}_i) \;\; = \;\; \frac{1}{n}\sum_{t=1}^{n} \text{Loss}^*\Big(y_t, \hat{f}_i(\mathbf{x}_t)\Big) \tag{1}$$

that we can compute to the (expected) risk $R(\hat{f}_i)$

$$R(\hat{f}_i) \;\; = \;\; E_{(\mathbf{x},y)\sim P}\Big\{ \text{Loss}^*\Big(y, \hat{f}_i(\mathbf{x})\Big) \Big\} \tag{2}$$

that we would like to have. In fact, we would like to keep these somewhat close so that the empirical risk (training error) still reflects how well the method will generalize. The empirical risk is computed on the basis of the available training set $S_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and the loss function $\text{Loss}^*(\cdot,\cdot)$ rather than say the hinge loss. For our purposes $\hat{f}_i \in \mathcal{F}_i$ could be any estimate derived from the training set that approximately tries to minimizing the empirical risk. In our analysis we will assume that $\text{Loss}^*(\cdot,\cdot)$ is the zero-one loss (classification error).

We'd like to quantify how much $R(\hat{f}_i)$ can deviate from $R_n(\hat{f}_i)$. The more powerful our set of classifiers is the more we would expect them to deviate from one another. In other words, the more choices we have in terms of discriminant functions, the less representative the training error of the minimizing classifier is about its generalization error. So, our goal is to show that

$$R(\hat{f}_i) \leq R_n(\hat{f}_i) + C(n, \mathcal{F}_i, \delta) \tag{3}$$

where the *complexity penalty* $C(n, \mathcal{F}_i)$ only depends on the model $\mathcal{F}_i$, the number of training instances, and a parameter δ. The peanalty does *not* depend on the actual training data. We will discuss the parameter δ below in more detail. For now, it suffices to say that $1 - \delta$

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

specifies the probability that the bound holds. We can only give a probabilistic guarantee in this sense since the empirical risk (training error) is a random quantity that depends on the specific instantiation of the data.

For nested models, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$, the penalty is necessarily an increasing function of i, the model order (e.g., the degree of polynomial kernel). Moreover, the penalty should go down as a function n. In other words, the more data we have, the more complex models we expect to be able to fit and still have the training error close to the generalization error.

The type of result in Eq.(3) gives us an *upper bound guarantee of generalization error*. We can then select the model with the best guarantee, i.e., the one with the lowest bound. Figure 1 shows how we would expect the upper bound to behave as a function of increasingly complex models in our nested "hierarchy" of models.

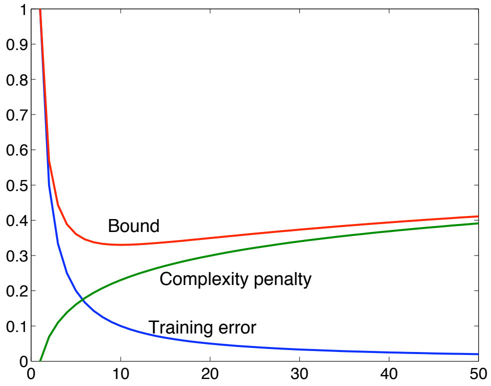

Figure 1: Bound on the generalization error as a function of model order (e.g., degree of polynomial kernel).

Let's derive a result of this type in the simple context where $\mathcal{F}_i$ only contains a finite number of classifiers $|\mathcal{F}_i| < \infty$. We will get to the general theory later on but this simple setting is helpful in understanding how such results come about. To avoid the question of how exactly we estimate $\hat{f}_i$, we will require a stronger guarantee: the bound should hold for *all* the classifiers in our set. Specifically, we try to find a tight upper bound on

$$P\left(\max_{f \in \mathcal{F}_i} |R(f) - R_n(f)| > \epsilon\right) \leq \delta \tag{4}$$

This is the probability that at least one classifier in our set deviates by more than ϵ from its training error. The probability is taken over the choice of the training data. So, if we

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

used ϵ to claim that

$$R(f) \leq R_n(f) + \epsilon \quad \text{for all } f \in \mathcal{F}_i \tag{5}$$

then this expression would fail with probability

$$\delta = P\left(\max_{f \in \mathcal{F}_i} |R(f) - R_n(f)| > \epsilon\right) \tag{6}$$

or, put another way, it would hold with probability $1 - \delta$ over the choice of the training data. If we fix δ, then the smallest $\epsilon = \epsilon(n, \mathcal{F}_i, \delta)$ that satisfies Eq.(6) is the complexity penalty we are after. Note that since the expression holds for all $f \in \mathcal{F}_i$ it necessarily also holds for $\hat{f}_i$.

In most cases we cannot compute δ exactly from Eq.(6) but we can derive an upper bound. This upper bound will lead to a larger than necessary complexity penalty but at least we will get a closed form expression (the utility of the model selection criterion will indeed depend on how tight a bound we can obtain). We will proceed as follows:

$$\begin{aligned} P\left(\max_{f \in \mathcal{F}_i} |R(f) - R_n(f)| > \epsilon\right) &= P\left(\exists f \in \mathcal{F}_i \text{ s.t. } |R(f) - R_n(f)| > \epsilon\right) && (7) \\ &\leq \sum_{f \in \mathcal{F}_i} P\left(|R(f) - R_n(f)| > \epsilon\right) && (8) \end{aligned}$$

where we have used the *union bound* $P(A_1 \cup A_2 \cup \ldots) \leq P(A_1) + P(A_2) + \ldots$ for any set of events $A_1, A_2, \ldots$ (not necessarily disjoint). In other words, we bound the probability that there are functions in our set with larger than ϵ deviation by a sum that each function individually has more than ϵ deviation between training and generalization errors.

Now, the discriminant function is fixed in any individual term in the sum

$$P\left(|R(f) - R_n(f)| > \epsilon\right) \tag{9}$$

It won't change as a function of the training data. We can then associate with each *i.i.d.* training sample $(\mathbf{x}_t, y_t)$, an indicator $s_t \in \{0, 1\}$ of whether the sample disagrees with f: $s_t = 1$ iff $y_t f(\mathbf{x}_t) \leq 0$. The empirical error $R_n(f)$ is therefore just an average of independent random variables (indicators) s_t:

$$R_n(f) = \frac{1}{n} \sum_{t=1}^{n} s_t \tag{10}$$

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

What is the expected value of each s_t when the expectation is taken over the choice of the training data? It's just $R(f)$, the expected risk. So, we can rewrite

$$P\left(|R(f) - R_n(f)| > \epsilon\right) \tag{11}$$

as

$$P\left(|q - \frac{1}{n}\sum_{t=1}^{n} s_t| > \epsilon\right) \tag{12}$$

where q equals $R(f)$ and the probability is now over n independent binary random variables $s_1, \ldots, s_n$ for which $P(s_t = 1) = q$. There are now standard results for evaluating a bound on how much an average of binary random variables deviates from its expectation (Hoeffding's inequality):

$$P\left(|q - \frac{1}{n}\sum_{t=1}^{n} s_t| > \epsilon\right) \leq 2\exp(-2n\epsilon^2) \tag{13}$$

Note that the bound does not depend on q (or $R(f)$) and therefore not on which f we chose. Using this result in Eq.(8), gives

$$P\left(\max_{f\in\mathcal{F}_i} |R(f) - R_n(f)| > \epsilon\right) \;\; \leq \;\; 2|\mathcal{F}_i|\exp(-2n\epsilon^2) = \delta \tag{14}$$

The last equality relates δ, $|\mathcal{F}_i|$, n, and ϵ, as desired. By solving for ϵ we get

$$\epsilon = \epsilon(n, \mathcal{F}_i, \delta) = \sqrt{\frac{\log|\mathcal{F}_i| + \log(2/\delta)}{2n}} \tag{15}$$

This is the complexity penalty we were after in this simple case with only a finite number of classifiers in our set.

We have now showed that with probability at least $1 - \delta$ over the choice of the training set,

$$R(f) \leq R_n(f) + \sqrt{\frac{\log|\mathcal{F}_i| + \log(2/\delta)}{2n}}, \quad \text{uniformly for all } f \in \mathcal{F}_i \tag{16}$$

So, for model selection, we would then estimate $\hat{f}_i \in \mathcal{F}_i$ for each model, plug the resulting $\hat{f}_i$ and $|\mathcal{F}_i|$ on the right hand side of the above equation, and choose the model with the lowest bound. n and δ would be the same for all models under consideration.

As an example of another way of using the result, suppose we set $\delta = 0.05$ and would like any classifier that achieves zero training error to have at most 10% generalization error. Let's solve for the number of training examples we would need for such a guarantee within model $\mathcal{F}_i$. We want

$$R(f) \leq 0 + \sqrt{\frac{\log|\mathcal{F}_i| + \log(2/0.05)}{2n}} \leq 0.10 \tag{17}$$

Solving for n gives

$$n = \frac{\log|\mathcal{F}_i| + \log(2/0.05)}{2(0.10)^2} \tag{18}$$

training examples.

Model selection criteria: Bayesian score, Bayesian information criterion

It is perhaps the easiest to explain the Bayesian score with an example. We will start by providing a Bayesian analysis of a simple linear regression problem. So, suppose our model $\mathcal{F}$ takes a $d-$dimensional input $\mathbf{x}$ and maps it to a real valued output y (a distribution over y) according to:

$$P(y|\mathbf{x}, \theta, \sigma^2) = N(y;\, \theta^T\mathbf{x}, \sigma^2) \tag{19}$$

where $N(y; \theta^T\mathbf{x}, \sigma^2)$ is a normal distribution with mean $\theta^T\mathbf{x}$ and variance σ^2. To keep our calculations simpler, we will keep σ^2 fixed and only try to estimate θ. Now, given any set of observed data $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, we can define the likelihood function

$$L(D;\theta) = \prod_{t=1}^{n} N(y_t;\, \theta^T\mathbf{x}_t,\, \sigma^2) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_t - \theta^T\mathbf{x}_t)^2\right) \tag{20}$$

We have previously used only the maximizing parameters $\hat{\theta}$ as estimates of the underlying parameter value (if any). In Bayesian analysis we are no longer satisfied with selecting a single linear regression function but would like to keep all of them, just weighted by their ability to explain the data, i.e., weighted by the corresponding likelihood $L(D;\theta)$. From this perspective, our knowledge about the parameter θ *after seeing the data* is defined by the *posterior distribution* $P(\theta|D)$ proportional to the likelihood

$$P(\theta|D) \propto L(D;\theta) \tag{21}$$

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

In many cases we cannot normalize this distribution, however. Suppose, as an extreme example, that we have no data. The likelihood function in this case is just one for all the parameter values. As a result the "posterior" after seeing no data is not well defined as a distribution (we cannot normalize the distribution by $\int 1\, d\theta = \infty$). To correct this problem it is advantageous to also put our prior belief about the parameter values in a form of a distribution, the *prior distribution* $P(\theta)$. This distribution captures what we believe about the parameter values before seeing any data. Similarly to the regularization penalty, we will typically choose the prior to prefer small parameter values, e.g.,

$$P(\theta) = N(\theta;\, 0,\, \sigma_p^2 \cdot I) \tag{22}$$

which is a zero mean spherical Gaussian (same variance in all directions). The smaller σ_p^2 is, the smaller values of θ we prefer prior to seeing the data. The posterior distribution, now well-defined as a distribution regardless of how much data we see, is proportional to the prior distribution $P(\theta)$ times the likelihood:

$$P(\theta|D) \propto L(D;\theta)P(\theta) \tag{23}$$

The normalization constant for the posterior, also known as the *marginal likelihood*, is given by

$$P(D|\mathcal{F}) = \int L(D;\theta)P(\theta)d\theta \tag{24}$$

and depends on the model $\mathcal{F}$ and the data but not specific parameter values. In our regression context, we can actually evaluate this marginal likelihood in closed form:

$$\begin{aligned}\log P(D|\mathcal{F}) &= -\frac{n}{2}\log(2\pi\sigma^2) + \frac{d}{2}\log\lambda - \frac{1}{2}\log|\mathbf{X}^T\mathbf{X} + \lambda I| &(25)\\ &\quad -\frac{1}{2\sigma^2}\left(\|\mathbf{y}\|^2 - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda I)^{-1}\mathbf{X}^T\mathbf{y}\right) &(26)\end{aligned}$$

where $\lambda = \sigma^2/\sigma_p^2$ (ratio of noise to prior variance), $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, and $\mathbf{y} = [y_1, \ldots, y_n]^T$. These definitions are identical to the regularized least squares regression discussed earlier.

The posterior distribution over the parameters is simply normalized by the marginal likelihood:

$$P(\theta|D) = \frac{L(D;\theta)P(\theta)}{P(D|\mathcal{F})} \tag{27}$$

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

In our context the posterior is also Gaussian $P(\theta|D) = N(\theta;\, \mu, \Sigma)$ with mean μ and covariance Σ given by

$$\begin{aligned} \mu &= (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} && (28) \\ \Sigma &= \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} && (29) \end{aligned}$$

Note that the posterior mean of the parameters is exactly the parameter estimate we derived earlier using penalized log-likelihood with the same prior. This is not an accident when all the distributions involved are indeed Gaussians. It is also worth pointing out that $P(\theta|D)$ is *very* different from the normal distribution over $\hat{\theta}$ we derived earlier when assuming that the responses $\mathbf{y}$ came from a linear model of the same type. We have made no such assumption here and the distribution $P(\theta|D)$ is defined on the basis of the single observed $\mathbf{y}$.

In Bayesian analysis the prediction of y in response to a new $\mathbf{x}$ would be given by weighting predictions based on individual θ's by the posterior distribution:

$$P(y|\mathbf{x}, D) = \int P(y|\mathbf{x}, \theta)P(\theta|D)d\theta \qquad (30)$$

So what is the model selection problem in this context? A true Bayesian would refrain from selecting a single model but include all of them in proportion to their ability to explain the data (just as with parameters). We will not go that far, however, but instead try to select different regression models, specified by different feature mappings $\mathbf{x} \rightarrow \phi(\mathbf{x})$. Let's consider then two regression models specified by linear $\phi^{(1)}(\mathbf{x})$ and quadratic $\phi^{(2)}(\mathbf{x})$ feature mappings. The models we compare are therefore

$$\begin{aligned} \mathcal{F}_1 : &\quad P(y|\mathbf{x}, \theta, \sigma^2) = N(y;\, \theta^T\phi^{(1)}(\mathbf{x}), \sigma^2),\;\; \theta \in \mathcal{R}^{d_1},\;\; P(\theta|\mathcal{F}_1) && (31) \\ \mathcal{F}_2 : &\quad P(y|\mathbf{x}, \theta, \sigma^2) = N(y;\, \theta^T\phi^{(2)}(\mathbf{x}), \sigma^2),\;\; \theta \in \mathcal{R}^{d_2},\;\; P(\theta|\mathcal{F}_2) && (32) \end{aligned}$$

Note that θ is of different dimension in the two models and thus the prior distributions over the parameters, $P(\theta|\mathcal{F}_1)$ and $P(\theta|\mathcal{F}_2)$, will have to be different. You might be wondering that since we are including the specification of the prior distribution as part of the model, the result will depend on how we selected the priors. Indeed, but not strongly so. This dependence on the prior is both an advantage and a disadvantage from the model selection point of view. We will discuss this further later on.

So, how do we select between the two competing models? We simply select the one whose marginal likelihood (Bayesian score[1]) is larger. In other words, after seeing data D we

[1] The definition of the Bayesian score often includes a prior over the models as well, e.g., how much we

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

would select model $\mathcal{F}_1$ if

$$P(D|\mathcal{F}_1) > P(D|\mathcal{F}_2) \tag{33}$$

Model selection criteria: Bayesian information criterion

Bayesian Information Criterion or BIC for short is an asymptotic approximation to the Bayesian score. It is frequently used for its simplicity. The criterion is simply

$$BIC = l(D;\hat{\theta}) - \frac{d}{2}\log(n) \tag{34}$$

where $l(D;\theta)$ is the log-likelihood of the data, $\hat{\theta}$ is the maximum likelihood estimate of the parameters, and d is the number of independent parameters in the model; n is the number of training examples as before. BIC is what the Bayesian score will converge to in the limit of large n. The Bayesian score is typically difficult to evaluate in practice and BIC serves as a simple tractable alternative. Similarly to the Bayesian score (marginal likelihood), we would select the model with the largest BIC score.

would prefer the simpler model before seeing any data. We have no reason to prefer one over another and therefore has used the same prior probability for both. As a result, the selection is carried out entirely on the basis of the marginal likelihood.

Cite as: Tommi Jaakkola, course materials for 6.867 Machine Learning, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].