

# Introduction to Statistical Learning Theory

## Lecture 7

In the last unit we looked at regularization - adding a  $\|w\|^2$  penalty.

We add a bias - we prefer classifiers with low norm.

How to incorporate more complicated prior knowledge?

Example: We trained many different face detectors  $w_1, \dots, w_k$  and have a probabilistic model for  $P(w)$ .

PAC-Bayes combines a Bayesian approach with an agnostic approach to analyse this situation.

We will start with an quick overview of Bayesian method.

Assume your data is drawn from a distribution that comes from some parametric family.

Example:  $P(y|x; w) = \mathcal{N}(w^T x, \sigma^2) = w^T x + \mathcal{N}(0, \sigma^2)$ . For simplicity we assume  $\sigma$  is a known fixed parameter.

Given a sample  $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$  we define the likelihood of  $w$  as

$$\mathcal{L}(w, S) = \log(P(y_1, \dots, y_m | x_1, \dots, x_m; w)) = \sum_{i=1}^m \log(P(y_i | x_i; w))$$

The maximum likelihood returns  $w = \arg \max \mathcal{L}(w, S)$

In our example  $P(y|x; w) = \mathcal{N}(w^T x, \sigma^2) = w^T x + \mathcal{N}(0, \sigma^2)$ .

This means that  $P(y_i|x_i; w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - w^T x_i)^2}{\sigma^2}\right)$ . We conclude that the likelihood is  $\mathcal{L}(w, S) = -\sum_{i=1}^m \frac{1}{\sigma^2} (y_i - w^T x_i)^2 + C$  where  $C$  is the normalization factors that do not depend on  $w$ .

In this model, maximum likelihood is equivalent to minimizing square loss.

Problem is - we want to maximize  $P(w|x, y)$ .

To get  $P(w|x, y)$  we need to a prior distribution  $P(w)$ .

We now have  $P(y|x, w)$  and  $P(w)$  so from Bayes theorem we get

$$P(w|x, y) = \frac{P(y|x, w) \cdot P(w)}{P(y|x)} \propto P(y|x, w) \cdot P(w)$$

The maximum a-posteriori (MAP) model is

$$w = \arg \max \{P(Y|X, w) \cdot P(w)\} = \arg \max \{\mathcal{L}(w, S) + \log(P(w))\}$$

Continuing our example - assume  $P(w) = \mathcal{N}(0, \sigma_w^2 \cdot I)$ .

We now get

$$\begin{aligned} w &= \arg \max \left[ - \sum_{i=1}^m \frac{1}{\sigma^2} (y_i - w^T x)^2 - \frac{1}{\sigma_w^2} \|w\|_2^2 \right] \\ &= \arg \min \left[ \sum_{i=1}^m (y_i - w^T x)^2 + \frac{\sigma^2}{\sigma_w^2} \|w\|_2^2 \right] \end{aligned}$$

This is equivalent to doing regularized ERM with  $\ell_2$  regularization. If we use Laplacian distribution instead, we will get  $\ell_1$  regularization.

MAP picks the best model, given our model and data. But why do we have to pick one model?

We have seen that the optimal classifier can be calculated given  $P(y|x)$  (assignment 1).

The Bayesian approach does exactly that, so we get

$$P(y|x, S) = \int_w P(y|x, w) \cdot P(w|S) dP(w)$$

Some cases (Gaussian) this has an analytic solution, most of the time there isn't any.

PAC-Bayes: We will consider algorithms that return a posterior - a distribution  $Q$  on  $\mathcal{H}$ .

### Definition 2.1 (Loss of posterior)

Let  $Q$  be a distribution on  $\mathcal{H}$ ,  $\mathcal{D}$  a distribution on  $\mathcal{X} \times \mathcal{Y}$  and  $S$  a finite sample. Define

$$L_{\mathcal{D}}(Q) = \mathbb{E}_{h \sim Q} [L_{\mathcal{D}}(h)] = \mathbb{E}_{h \sim Q} \left[ \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)] \right]$$

$$L_S(Q) = \mathbb{E}_{h \sim Q} [L_S(h)] = \mathbb{E}_{h \sim Q} \left[ \frac{1}{m} \sum_{i=1}^m \ell(h, z_i) \right]$$



We can turn a posterior into a learning algorithm:

### Definition 2.2 (Gibbs hypothesis)

Let  $Q$  be a distribution on  $\mathcal{H}$ . The Gibbs hypothesis is the following randomized hypothesis - Given  $x$ , sample  $h$  according to  $Q$  and return  $h(x)$ .

It is straightforward to show that the expected loss is  $L_{\mathcal{D}}(Q)$ .

We want to show that if  $Q$  is similar to  $P$  we generalize well.  
Kullback-Leibler (KL) divergence is how we measure similarity.

### Definition 2.3 (KL Divergence)

Let  $P, Q$  be continuous or discrete distributions. Define

$$KL(Q||P) = \mathbb{E}_{x \sim Q} \left[ \ln \left( \frac{Q(x)}{P(x)} \right) \right]$$

Notice this is not symmetrical  $KL(Q||P) \neq KL(P||Q)$ .

The intuition behind this definition comes from information theory.

## KL Divergence

Assume we have a finite alphabet and message  $x$  is sent with probability  $P(x)$ .

Shannon's coding theorem states that if you code  $x$  with  $\log_2(1/P(x))$  bits you get an optimal coding. The expected bits per letter is then

$$\mathbb{E}_{x \sim P} \left[ \log_2 \left( \frac{1}{P(x)} \right) \right] = H(P).$$

Consider now that we use the optimal code for  $P$ , but the letters are sent according to  $Q$ . The expected bits per letter is now

$$\mathbb{E}_{x \sim Q} \left[ \log_2 \left( \frac{1}{P(x)} \right) \right] = \mathbb{E}_{x \sim Q} \left[ \log_2 \left( \frac{Q(x)}{P(x)} \right) + \log_2 \left( \frac{1}{Q(x)} \right) \right] = H(Q) + KL(Q||P)$$

Up to a factor due to different log basis. This shows  $KL(Q||P) \geq 0$ .

Another perspective - The mutual information  $I(X, Y)$  is equal

$$I(X, Y) = KL(P(X, Y) || P(X)P(Y)).$$

Example 1:  $P$  some distribution on  $x_1, \dots, x_m$ ,  $Q$  is 1 on  $x_i$  then  $KL(Q||P) = \ln(1/P(x_i))$ .

Example 2: If  $P(x_i) = 0$  and  $Q(x_i) > 0$  then  $KL(Q||P) = \infty$ .

Example 3: If  $\alpha, \beta \in [0, 1]$  then  $KL(\alpha||\beta) \equiv KL(\text{Bernoulli}(\alpha)||\text{Bernoulli}(\beta)) = \alpha \ln\left(\frac{\alpha}{\beta}\right) + (1 - \alpha) \ln\left(\frac{1-\alpha}{1-\beta}\right)$

Example 4: If  $Q = \mathcal{N}(\mu_0, \Sigma_0)$  and  $P = \mathcal{N}(\mu_1, \Sigma_1)$  Gaussian distributions in dimension  $n$ , then

$$KL(Q||P) = \frac{1}{2} \left( \text{trace}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)\Sigma_1^{-1}(\mu_1 - \mu_0) - n - \frac{\det \Sigma_0}{\det \Sigma_1} \right)$$

We will now prove the following bound:

### Theorem 3.1 (McAllester)

*Let  $Q, P$  be distributions on  $\mathcal{H}$  and  $\mathcal{D}$  be a distribution on  $\mathcal{X} \times \mathcal{Y}$ . Assume  $\ell(h, z) \in [0, 1]$ . Let  $S \sim \mathcal{D}^m$  be a sample, then with probability greater or equal to  $1 - \delta$  over  $S$  we have*

$$KL(L_S(Q) || L_{\mathcal{D}}(Q)) \leq \frac{KL(Q || P) + \ln\left(\frac{m+1}{\delta}\right)}{m} \quad (1)$$

Notice: that the l.h.s is the KL divergence between two numbers (as in example 3), while the r.h.s is between distributions.

Also notice we assume no connection between  $\mathcal{D}$  and  $P$  - it is still an agnostic analysis.

We will split the proof into technical lemmas:

### Lemma 3.1

*If  $X$  is a real valued random number satisfying  $P(X \leq x) \leq e^{-mf(x)}$ , then following holds:  $\mathbb{E}[e^{(m-1)f(x)}] \leq m$ .*

Proof: Define  $F(x) = P(X \leq x)$  the CDF then from basic properties of the CDF we have  $P(F(x) \leq y) \leq y$ , therefore  $P(e^{-mf(x)} \leq y) \leq y$ . So

$$y \geq P(e^{-mf(x)} \leq y) = P(e^{mf(x)} \geq 1/y) = P\left(e^{(m-1)f(x)} \geq y^{-\frac{m-1}{m}}\right) \quad (2)$$

Define  $\nu = y^{-\frac{m-1}{m}}$  and we have  $P(e^{(m-1)f(x)} \geq \nu) \leq \nu^{\frac{-m}{m-1}}$ .

We use the following fact: for non-negative r.v we have

$$\mathbb{E}[W] = \int_0^{\infty} P(W \geq \nu) d\nu.$$

We conclude:

$$\begin{aligned}\mathbb{E}[e^{(m-1)f(x)}] &= \int_0^\infty P(e^{(m-1)f(x)} \geq \nu) d\nu \leq 1 + \int_1^\infty \nu^{\frac{-m}{m-1}} d\nu \\ &= 1 - (m-1) \left[ \nu^{-1/(m-1)} \right]_1^\infty = m\end{aligned}$$

□

We will use the stronger version of the Hoeffding bound we proved in Lecture 1:

### Lemma 3.2 (Hoeffding)

If  $X_1, \dots, X_m$  are i.i.d r.v such that  $X_i \in [0, 1]$ , and  $\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$  then for  $\epsilon \in [0, 1]$  we have the following

$$P(\bar{X} \leq \epsilon) \leq e^{-mKL(\epsilon || \mathbb{E}[X_1])}$$

## Lemma 3.3

With probability greater than  $1 - \delta$  over  $S$ ,

$$\mathbb{E}_{h \sim P} \left[ e^{(m-1)KL(L_S(h) \| L_{\mathcal{D}}(h))} \right] \leq \frac{m}{\delta}$$

Proof sketch - using lemma 3.1 + 3.2 (Hoeffding) we get that for any  $h \in \mathcal{H}$  we have  $\mathbb{E}_{S \sim \mathcal{D}^m} \left[ e^{(m-1)KL(L_S(h) \| L_{\mathcal{D}}(h))} \right] \leq m$ . The lemma follows by taking expectation w.r.t  $P$  and Markov's inequality.

Finally we need this shift of measure theorem:

## Lemma 3.4

$$\mathbb{E}_{x \sim Q} [f(x)] \leq KL(Q \| P) + \ln \mathbb{E}_{x \sim P} [e^{f(x)}]$$



Proof:

$$\begin{aligned}
 \mathbb{E}_{x \sim Q} [f(x)] &= \mathbb{E}_{x \sim Q} \left[ \ln e^{f(x)} \right] = \mathbb{E}_{x \sim Q} \left[ \ln \left( \frac{P(x)}{Q(x)} e^{f(x)} \right) + \ln \frac{Q(x)}{P(x)} \right] \\
 &= KL(Q||P) + \mathbb{E}_{x \sim Q} \left[ \ln \left( \frac{P(x)}{Q(x)} e^{f(x)} \right) \right] \\
 &\leq KL(Q||P) + \ln \left( \mathbb{E}_{x \sim Q} \left[ \frac{P(x)}{Q(x)} e^{f(x)} \right] \right) \\
 &= KL(Q||P) + \ln \left( \mathbb{E}_{x \sim P} \left[ e^{f(x)} \right] \right)
 \end{aligned}$$

where we use Jensen's inequality.

We Can now prove theorem 3.1:

### Theorem 3.1 (McAllester)

*Let  $Q, P$  be distributions on  $\mathcal{H}$  and  $\mathcal{D}$  be a distribution on  $\mathcal{X} \times \mathcal{Y}$ . Assume  $\ell(h, z) \in [0, 1]$ . Let  $S \sim \mathcal{D}^m$  be a sample, then with probability greater or equal to  $1 - \delta$  over  $S$  we have*

$$KL(L_S(Q) || L_{\mathcal{D}}(Q)) \leq \frac{KL(Q || P) + \ln\left(\frac{m+1}{\delta}\right)}{m} \quad (1)$$

Proof: Define  $f(h) = KL((L_S(h)||L_D(h)))$ . Using the shift of measure (lemma 3.4) and lemma 3.3 we get:

$$\mathbb{E}_{h \sim Q}[mf(h)] \leq KL(Q||P) + \ln \mathbb{E}_{h \sim P}[e^{mf(h)}] \leq KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)$$

With probability greater or equal to  $1 - \delta$ .

To finish the proof we will use the fact that  $KL$  divergence is convex, so from the Jensen inequality

$$\begin{aligned} KL(L_S(Q)||L_D(Q)) &= KL(\mathbb{E}_Q[L_S(h)]||\mathbb{E}_Q[L_D(h)]) \\ &\leq \mathbb{E}_Q[KL((L_S(h)||L_D(h)))] = \mathbb{E}_Q[f(h)]. \end{aligned}$$

□

(sweeping a few subtleties under the rug)

We bounded  $KL(L_S(Q)||L_{\mathcal{D}}(Q))$ . Next step - bound  $L_{\mathcal{D}}(Q) - L_S(Q)$ .  
We will show two bounds using the following lemma:

### Lemma 3.5

*If  $a, b \in [0, 1]$  and  $KL(a||b) \leq x$ , then  $b \leq a + \sqrt{\frac{x}{2}}$  and  $b \leq a + 2x + \sqrt{2ax}$*

Where the second is much stronger if  $a$ , i.e.  $L_S(Q)$  is very small.

## Generalization Bounds

Proof of first inequality: Fix  $b$  and define  $f(a) = KL(a||b) - 2(b - a)^2$ . The first and second derivatives are:

$$f'(a) = \ln\left(\frac{a}{1-a}\right) - \ln\left(\frac{b}{1-b}\right) - 4(a - b)$$

$$f''(a) = \frac{1}{a(1-a)} - 4$$

The function  $a(1-a)$  has its maximum at  $a = 1/2$  with value  $1/4$  so  $f''(a) \geq 0$ . As  $f'(b) = 0$  we have  $f(a)$  has its minimum at  $a = b$  with  $f(b) = 0$ .

Therefore  $2(a - b)^2 \leq KL(a||b) \leq x$  proving  $b \leq a + \sqrt{\frac{x}{2}}$ .

Second inequality is left as an exercise. □

Notice we also have  $b \geq a - \sqrt{\frac{x}{2}}$ .

We can combine everything to get the following theorem:

### Theorem 3.6 (Generalization Bound)

*Let  $Q, P$  be distributions on  $\mathcal{H}$  and  $\mathcal{D}$  be a distribution on  $\mathcal{X} \times \mathcal{Y}$ . Assume  $\ell(h, z) \in [0, 1]$ . Let  $S \sim \mathcal{D}^m$  be a sample, then with probability greater or equal to  $1 - \delta$  over  $S$  we have*

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{2m}}$$

$$\begin{aligned} L_{\mathcal{D}}(Q) &\leq L_S(Q) + 2 \frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{m} \\ &\quad + \sqrt{2L_S(Q) \frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{m}} \end{aligned}$$