Expectation-Prior PAC-Bayes Bounds for SVMs

Shiliang Sun

Department of Computer Science
University College London
shiliangsun@gmail.com

March 22, 2010

Joint work with John Shawe-Taylor and Emilio Parrado-Hernández

Roadmap

1. Expectation-Prior PAC-Bayes Bounds for SVMs
   - Single-Expectation-Prior PAC-Bayes Bound
   - Multiple-Expectation-Prior PAC-Bayes Bound
   - Expectation-Prior PAC-Bayes Bound with Non-identity Gaussian
2. Insights for Training SVMs
3. Some Other Research and Experimental Results (by Emilio)

Expectation-Prior Based Bound: Motivation & Setting

Motivation:

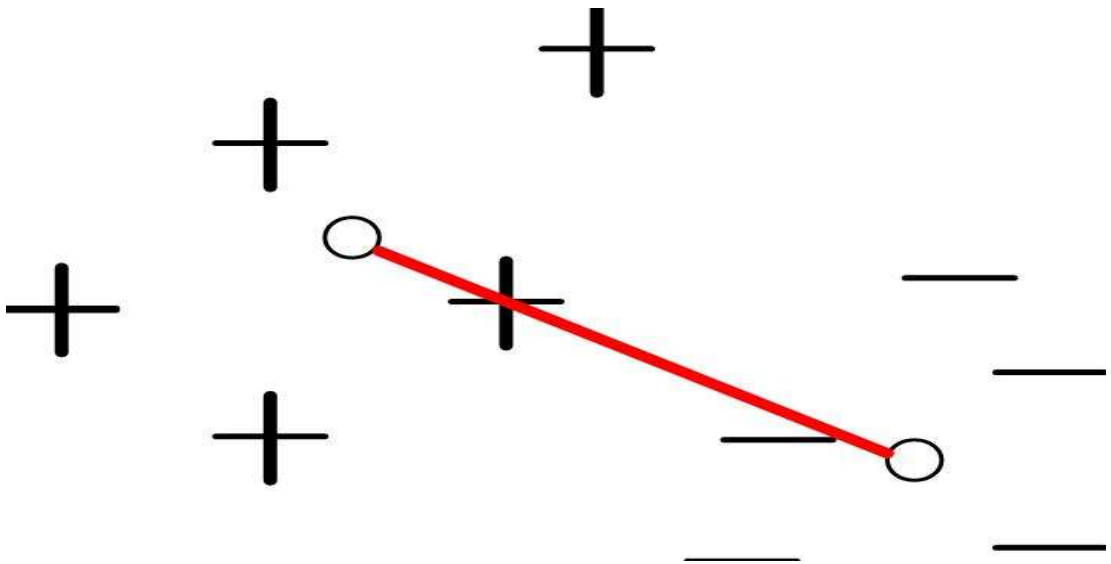
- To design a good prior without seeing any data.
  'Good' here means potentially tighter generalization bounds.

Setting:

- Binary classification with labels in $\{+1, -1\}$
- linear threshold classifier:
  $c(x) = \text{sign}(\mathbf{w}^\top \phi(x))$

Single-Expectation-Prior PAC-Bayes Bound

- Define $\mathbf{w}_p \triangleq E_{(x,y)\sim\mathcal{D}}[y\phi(x)]$ where $y \in \{+1, -1\}$. We use this general expectation form in deriving bounds.
- A special case of $\mathbf{w}_p$ is $\frac{1}{2}(\mathbf{w}^+ - \mathbf{w}^-)$ with $\mathbf{w}^+ \triangleq E_{(x,y)\sim\mathcal{D},y=+1}[\phi(x)]$, $\mathbf{w}^- \triangleq E_{(x,y)\sim\mathcal{D},y=-1}[\phi(x)]$ when each class has the same prior probability $1/2$.



Single-Expectation-Prior PAC-Bayes Bound

- The expectation cannot be exactly evaluated given limited information from the training data. Using its empirical estimate to approximate it!
- Given a sample set S including m examples, the empirical estimate of $\mathbf{w}_p$ would be

$$\hat{\mathbf{w}}_p = E_{(x,y)\sim S}[y\phi(x)] = \frac{1}{m}\sum_{i=1}^{m}[y_i\phi(x_i)].$$

Single-Expectation-Prior PAC-Bayes Bound

Theorem

For all $\mathcal{D}$, for all Gaussian prior $P \sim \mathcal{N}(\eta \mathbf{w}_p, I)$ over margin classifiers, for all $\delta \in (0, 1]$:

$$Pr_{S\sim\mathcal{D}^m}(\forall \mathbf{w}, \mu : KL_+(\hat{Q}_S(\mathbf{w}, \mu)||Q_D(\mathbf{w}, \mu)) \leq \frac{\frac{1}{2}(\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\| + \eta\frac{R}{\sqrt{m}}(2 + \sqrt{2\ln\frac{2}{\delta}}))^2 + \ln(\frac{2(m+1)}{\delta})}{m}) \geq 1 - \delta,$$

where the posterior is $Q \sim \mathcal{N}(\mu\tilde{\mathbf{w}}, I)$ with $\tilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$, and $R = \sup_x \|\phi(x)\|$.

- R is ready to compute for some kernels, e.g., $R = 1$ for $k(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2/(2\sigma^2))$.

Proof.

First, we try to bound $KL(Q||P)$. We have

$$
\begin{aligned}
& KL(Q||P) \\
=& \frac{1}{2}\|\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p\|^2 \\
=& \frac{1}{2}\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p + \eta\hat{\mathbf{w}}_p - \eta\mathbf{w}_p\|^2 \\
=& \frac{1}{2}\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|^2 + \frac{1}{2}\|\eta\hat{\mathbf{w}}_p - \eta\mathbf{w}_p\|^2 + (\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p)^\top(\eta\hat{\mathbf{w}}_p - \eta\mathbf{w}_p) \\
\leq& \frac{1}{2}\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|^2 + \frac{1}{2}\eta^2\|\hat{\mathbf{w}}_p - \mathbf{w}_p\|^2 + \eta\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|\|\hat{\mathbf{w}}_p - \mathbf{w}_p\|. \quad (1)
\end{aligned}
$$

where the last inequality uses Cauchy-Schwarz inequality. Now it suffices to bound $\|\hat{\mathbf{w}}_p - \mathbf{w}_p\|$.

Proof. (cont.)

Define $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$. It is simple to show that $\sup_{(\mathbf{x},y)} \|y\phi(\mathbf{x})\| = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\| = R$. With reference to a result on estimating the center of mass, we have

$$Pr\left(\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| \geq \frac{2R}{\sqrt{m}} + \epsilon\right) \leq \exp(-\frac{2m\epsilon^2}{4R^2}). \tag{2}$$

Setting the right hand side equal to $\delta/2$, solving for ϵ shows that with probability at least $1 - \delta/2$, we have

$$\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| \leq \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right). \tag{3}$$

Proof. (cont.)

Define $b = \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right)$ and $a = \|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|$, we have

$$Pr_{S\sim\mathcal{D}^m}\left(KL(Q||P) \leq \frac{1}{2}a^2 + \frac{1}{2}\eta^2 b^2 + \eta ab\right) \geq 1 - \delta/2. \tag{4}$$

Then, according to the general theorem on PAC-Bayes bound, we have

$$Pr_{S\sim\mathcal{D}^m}\left(\forall Q(c) : KL_+(\hat{Q}_S||Q_\mathcal{D}) \leq \frac{KL(Q||P) + \ln(\frac{2(m+1)}{\delta})}{m}\right) \geq 1 - \delta/2. \tag{5}$$

Combining (4) and (5) and using $(1-\delta/2)^2 > 1-\delta$, we get

$$Pr_{S\sim\mathcal{D}^m}\left(\forall \mathbf{w}, \mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_\mathcal{D}(\mathbf{w},\mu)) \leq \frac{\frac{1}{2}(a+\eta b)^2 + \ln(\frac{2(m+1)}{\delta})}{m}\right)$$

$\geq 1 - \delta$. This completes the proof. □

Multiple-Expectation-Prior PAC-Bayes Bound

The prior can be selected from a couple of candidates at the cost of an additional penalty. By applying the union bound, we get

Theorem

For all $\mathcal{D}$, for all Gaussian prior $P_j \sim \mathcal{N}(\eta_j \mathbf{w}_p, I)$ $(j = 1, \ldots, J)$ over margin classifiers (suppose each prior can be selected with positive weights $\{\pi_j\}_{j=1}^J$ and $\sum_{j=1}^J \pi_j = 1$), for all $\delta \in (0, 1]$:

$$Pr_{S\sim\mathcal{D}^m}(\forall \mathbf{w}, \mu : KL_+(\hat{Q}_S(\mathbf{w}, \mu)||Q_{\mathcal{D}}(\mathbf{w}, \mu)) \leq \min_j \frac{\frac{1}{2}(\|\mu\tilde{\mathbf{w}} - \eta_j \hat{\mathbf{w}}_p\| + \eta_j \frac{R}{\sqrt{m}}(2 + \sqrt{2 \ln \frac{2}{\delta}}))^2 + \ln(\frac{2(m+1)}{\delta}) + \ln \frac{1}{\pi_j}}{m}) \geq 1 - \delta,$$

where the posterior is $Q \sim \mathcal{N}(\mu\tilde{\mathbf{w}}, I)$ with $\tilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$, and $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$. For a uniform distribution over priors, $\ln \frac{1}{\pi_j} = \ln J$.

Expectation-Prior PAC-Bayes Bound with Non-identity Gaussian

Theorem

Consider a prior distribution $P \sim \mathcal{N}(\eta \mathbf{w}_p, I, \tau^2)$ *of classifiers consisting a Gaussian distribution centered on* $\eta \mathbf{w}_p$*, with identity covariance in all directions except* $\mathbf{w}_p$ *in which the variance is* τ^2*. Then, for all distributions* $\mathcal{D}$*, for all* $\delta \in (0, 1]$*, we have*

$$Pr_{S \sim \mathcal{D}^m}(\forall \mathbf{w}, \mu : KL_+(\hat{Q}_S(\mathbf{w}, \mu) || Q_{\mathcal{D}}(\mathbf{w}, \mu)) \leq$$
$$\frac{\frac{1}{2}(\ln(\tau^2) + \frac{(\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\| + \eta\frac{R}{\sqrt{m}}(2+\sqrt{2\ln\frac{2}{\delta}}))^2 - \mu^2 + 1}{\tau^2} + \mu^2 - 1) + \ln(\frac{2(m+1)}{\delta})}{m})$$
$$\geq 1 - \delta,$$

where the posterior is $Q \sim \mathcal{N}(\mu\tilde{\mathbf{w}}, I)$ *with* $\tilde{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$*, and* $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$*.*

Proof.

$$KL(Q||P) = \frac{1}{2}\left(\ln(\tau^2) + \frac{1}{\tau^2} - 1 + \frac{P^{\|}_{\mathbf{w}_p^*}(\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p)^2}{\tau^2} + P^{\perp}_{\mathbf{w}_p^*}(\mu\tilde{\mathbf{w}})^2\right) ,$$

with $\mathbf{w}_p^* = \mathbf{w}_p/\|\mathbf{w}_p\|$. $\frac{1}{\tau^2}P^{\|}_{\mathbf{w}_p^*}(\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p)^2 + P^{\perp}_{\mathbf{w}_p^*}(\mu\tilde{\mathbf{w}})^2$ can be rewritten as

$$\begin{aligned}
&\frac{1}{\tau^2}\left(\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}(\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p)\right)^2 + \|\mu\tilde{\mathbf{w}}\|^2 - \left(\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}\mu\tilde{\mathbf{w}}\right)^2 \\
=&\frac{1}{\tau^2}\left(\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}\mu\tilde{\mathbf{w}} - \eta\|\mathbf{w}_p\|\right)^2 + \|\mu\tilde{\mathbf{w}}\|^2 - \left(\frac{\mathbf{w}_p^\top}{\|\mathbf{w}_p\|}\mu\tilde{\mathbf{w}}\right)^2 \\
=&\frac{1}{\tau^2}(\eta^2\|\mathbf{w}_p\|^2 - 2\eta\mathbf{w}_p^\top\mu\tilde{\mathbf{w}}) + \|\mu\tilde{\mathbf{w}}\|^2 \\
=&\frac{1}{\tau^2}(\|\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p\|^2 - \|\mu\tilde{\mathbf{w}}\|^2) + \|\mu\tilde{\mathbf{w}}\|^2 \\
=&\frac{1}{\tau^2}(\|\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p\|^2 - \mu^2) + \mu^2 \, . \qquad (6)
\end{aligned}$$

Proof. (cont.)

By equation (1), we have

$$\|\mu\tilde{\mathbf{w}} - \eta\mathbf{w}_p\|^2 \leq \|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|^2 + \eta^2\|\hat{\mathbf{w}}_p - \mathbf{w}_p\|^2 + 2\eta\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| . \tag{7}$$

By equation (3), we have with probability at least $1 - \delta/2$

$$\|\hat{\mathbf{w}}_p - \mathbf{w}_p\| \leq \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right) . \tag{8}$$

With $a = \|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|$ and $b = \frac{R}{\sqrt{m}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right)$, we have

$$Pr_{S\sim\mathcal{D}^m}(KL(Q||P) \leq \frac{1}{2}(\ln(\tau^2) + \frac{1}{\tau^2} - 1 + \frac{a^2 + \eta^2 b^2 + 2\eta ab - \mu^2}{\tau^2} + \mu^2))$$
$$\geq 1 - \delta/2. \tag{9}$$

Proof. (cont.)

Combining (9) with

$$Pr_{S\sim\mathcal{D}^m}(\forall Q(c) : KL_+(\hat{Q}_S||Q_\mathcal{D}) \leq \frac{KL(Q||P) + \ln(\frac{2(m+1)}{\delta})}{m}) \geq 1-\delta/2 \tag{10}$$

results in

$$Pr_{S\sim\mathcal{D}^m}(\forall \mathbf{w}, \mu : KL_+(\hat{Q}_S(\mathbf{w},\mu)||Q_\mathcal{D}(\mathbf{w},\mu)) \leq \frac{\frac{1}{2}(\ln(\tau^2) + \frac{(a+\eta b)^2-\mu^2+1}{\tau^2} + \mu^2 - 1) + \ln(\frac{2(m+1)}{\delta})}{m}) \geq 1-\delta,$$

which completes the proof. □

Insights for Training SVMs

- A common term in the above expectation-prior PAC-Bayes bound is $\|\mu\tilde{\mathbf{w}} - \eta\hat{\mathbf{w}}_p\|$. It can be used to train SVMs to get a tighter generalization bound.
- For example, simply replace $\|\mathbf{w}\|$ from the objective function of SVMs with $\|\mathbf{w} - \eta\hat{\mathbf{w}}_p\|$.

Thank You!