CS281B/Stat241B. Statistical Learning Theory. Lecture 24.

Peter Bartlett

Overview

- Kernel regression.
  - Kernel ridge regression.
- Convex losses for classification.
  - Classification calibration.
  - Excess risk versus excess ϕ-risk.

Kernel methods for regression

Consider a regression problem:

- Probability distribution P on $\mathcal{X} \times \mathbb{R}$,
- Observe $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$,
- Choose $f_n : \mathcal{X} \to \mathbb{R}$ to minimize $\mathbb{E}\ell(Y, f(X))$ for $(X, Y) \sim P$.

Examples:

1. $\ell(y, \hat{y}) = (y - \hat{y})^2$.
2. $\ell(y, \hat{y}) = |y - \hat{y}|$.
3. $\ell(y, \hat{y}) = (|y - \hat{y}| - \epsilon)_+$.
   (ϵ-insensitive loss: gives a similar QP to the SVM)

Kernel ridge regression

For quadratic loss, $\ell(y, \hat{y}) = (y - \hat{y})^2$, we have

$$\min_{f \in \mathcal{H}} \qquad \lambda \|f\|_H^2 + \sum_{i=1}^{n} (y_i - f(x_i))^2 .$$

Choosing the slack variable ξ_i and introducing an equality constraint, we have

$$\begin{aligned} \min_{\theta, \xi} \qquad & \lambda \|\theta\|^2 + \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t.} \qquad & \xi_i = y_i - \theta^T x_i . \end{aligned}$$

Kernel ridge regression

Forming the Lagrangian (for an equality, we do not need a sign constraint on the dual variable) and eliminating the primal variables, we obtain:

$$\theta = \frac{1}{2\lambda} \sum \alpha_i x_i,$$
$$\xi_i = \frac{\alpha_i}{2},$$
$$g(\alpha) = y^T \alpha - \frac{1}{4\lambda} \alpha' K \alpha - \frac{1}{4} \alpha^T \alpha.$$

The solution to the dual problem is

$$\alpha = 2\lambda (K + \lambda I)^{-1} y.$$

Kernel ridge regression

This has a natural interpretation as a Bayesian method. The prediction rule $f_n(x)$ is the mean of the posterior distribution of $f(x)$ when $f : \mathcal{X} \to \mathbb{R}$ has a Gaussian process prior with $\mathbb{E}f(x_i) = 0$, $\mathrm{Var}(f(x_1), f(x_2)) = k(x_1, x_2)$, and $y = f(x) + \mathcal{N}(0, \lambda)$.

Convex loss for classification

We have seen various examples of convex loss functions used for classification. While we might aim to choose a decision rule $f : \mathcal{X} \to \mathbb{R}$ to minimize

$$R(f) = \Pr(Y \neq \text{sign}(f(X))) = \mathbb{E}1[Yf(X) \leq 0],$$

we often work with f chosen to minimize a (regularized version of a) sample average of a convex loss function like:

$$\begin{aligned}
\phi_{svm}(yf(x)) &= (1 - yf(x))_+ \,, \\
\phi_{AdaBoost}(yf(x)) &= \exp(-yf(x)), \\
\phi_{logistic}(yf(x)) &= \log(1 + \exp(-yf(x))).
\end{aligned}$$

This allows the use of efficient convex optimization algorithms. What is the cost of this computational convenience?

Convex loss for classification

We will ignore the issue of $\mathbb{E}\phi(Yf(X))$ versus $\hat{\mathbb{E}}\phi(Yf(X))$: suppose that we choose $f : \mathcal{X} \to \mathbb{R}$ to minimize $\mathbb{E}\phi(Yf(X))$. When does this lead to a good classifier (that is, with small risk)?

Define

$$\begin{aligned} \ell(y, f(x)) &= 1[yf(x) \le 0], \\ R(f) &= \mathbb{E}\ell(Y, f(X)), \\ R_\phi(f) &= \mathbb{E}\phi(Yf(X)). \\ \text{e.g., } \phi(yf(x)) &= (1 - yf(x))_+. \end{aligned}$$

First, we can observe that $\phi(yf(x)) \ge \ell(y, f(x))$ implies that $R(f) \le R_\phi(f)$. So a small $R_\phi(f)$ gives small $R(f)$. But this is a rather weak assurance if, for example, $\inf_f R_\phi(f) > 0$. When does minimizing R_ϕ lead to minimal R?

Convex loss for classification

Consider a *fixed* $x \in \mathcal{X}$.

$$
\begin{aligned}
\text{Define } \eta(x) &= \Pr(Y = 1 | X = x). \\
\text{Then } R_\phi(f) &= \mathbb{E}\phi(Y f(X)) \\
&= \mathbb{E}\mathbb{E}\left[\phi(Y f(X)) | X\right], \\
\mathbb{E}\left[\phi(Y f(X)) | X = x\right] &= \Pr(Y = 1 | X = x)\phi(f(x)) \\
&\qquad + \Pr(Y = -1 | X = x)\phi(-f(x)) \\
&= \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).
\end{aligned}
$$

Define the optimizer of this conditional expectation:

$$
H(\eta) := \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)\right)
$$

Examples

For $\phi(\alpha) = (1-\alpha)_+$,

$$H(\eta) = 2\min(\eta, 1-\eta),$$

$$H^-(\eta) = \phi(0) = 1,$$

$$\psi(\theta) = 1 - 2\min\left(\frac{1+\theta}{2}, \frac{1-\theta}{2}\right) = \theta.$$

Examples

For $\phi(\alpha) = \exp(-\alpha)$,

$$\begin{aligned} H(\eta) &= 2\sqrt{\eta(1-\eta)}, \\ H^-(\eta) &= \phi(0) = 1, \\ \psi(\theta) &= 1 - \sqrt{1-\theta^2}. \end{aligned}$$

Classification calibration

The prediction $\hat{y}$ with minimal conditional risk is $\mathrm{sign}(2\eta(x) - 1)$. If the optimal conditional expectation $\mathbb{E}[\phi(Yf(X))|X = x]$ can be achieved with a value of α with the wrong sign, then minimizing R_ϕ is not useful for classification. So define

$$H^-(\eta) := \inf \left\{ \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha) : \alpha(2\eta - 1) \le 0 \right\}.$$

Definition: We say that ϕ is **classification-calibrated** if, for all $\eta \neq 1/2$, $H^-(\eta) > H(\eta)$.

Classification-calibration is clearly necessary for minimization of R_ϕ to lead to minimization of R. We shall see that it is also sufficient.

Classification calibration for convex ϕ

Theorem: For ϕ convex, ϕ is classification-calibrated iff

1. ϕ is differentiable at 0,
2. $\phi'(0) < 0$.

Proof: *If* is straightforward to check.

Only if: suppose that ϕ is not differentiable at 0. Then convexity implies that it lies above several tangent lines. But then for values of η near $1/2$, $\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)$ is minimized by $\alpha = 0$, so ϕ is not classification-calibrated.

Also, $\phi'(0) \geq 0$ leads to $\mathrm{sign}(\alpha^*(\eta)) \neq \mathrm{sign}(\eta - 1/2)$.

Excess risk versus excess ϕ-risk

Theorem: For any nonnegative ϕ, measurable $f : \mathcal{X} \to \mathbb{R}$ and probability distribution P on $\mathcal{X} \times \{\pm 1\}$,

$$\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*,$$

where $R_\phi^* := \inf_f R_\phi(f)$, $R^* := \inf_f R(f)$, and, if ϕ is convex,

$$\psi(\theta) := H^- \left(\frac{1+\theta}{2} \right) - H \left(\frac{1+\theta}{2} \right)$$

Furthermore, ϕ is classification calibrated iff

$$\psi(\theta_i) \to 0 \text{ iff } \theta_i \to 0.$$

And if ϕ is classification calibrated and convex, $\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right)$.

Excess risk versus excess ϕ-risk

If ϕ is not convex, the theorem holds with $\psi = \tilde{\psi}^{**}$, the Legendre biconjugate of

$$\tilde{\psi}(\theta) := H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right).$$

(The biconjugate g^{**} of g is the largest convex lower bound on $\tilde{\psi}$, defined by $\operatorname{epi} g^{**} = \overline{\mathrm{co}} \operatorname{epi} g$. So the definitions are equivalent if ϕ is convex.)

Excess risk versus excess ϕ-risk: Proof

First, some observations about H and ψ:

1. $H(\eta) = H(1-\eta)$; $H^-(\eta) = H^-(1-\eta)$.
2. H is concave, ψ is convex.
3. $\psi(0) = 0$.
4. $\mathbb{E}H(\eta(X)) = R_\phi^*$.

Excess risk versus excess ϕ-risk: Proof

In Lecture 2, we saw that

$$R(f) - R^* = \mathbb{E}\left(1\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}\left(\eta(X) - \frac{1}{2}\right)\right]|2\eta(X) - 1|\right).$$

Since ψ is convex, Jensen's inequality implies

$$\begin{aligned}
\psi\left(R(f) - R^*\right) &\leq \mathbb{E}\psi\left(1\left[\cdots\right]|2\eta(X) - 1|\right) \\
&= \mathbb{E}1\left[\cdots\right]\psi\left(|2\eta(X) - 1|\right) && \text{(since } \psi(0) = 0\text{)} \\
&= \mathbb{E}1\left[\cdots\right]\left(H^-(\eta(X)) - H(\eta(X))\right) && \text{(def of } \psi\text{)}
\end{aligned}$$

Excess risk versus excess ϕ-risk: Proof

Now, $H^-(\eta(X))$ is the minimizer of $\mathbb{E}[\phi(Y\alpha)|X]$ when $\mathrm{sign}(\alpha) \neq \mathrm{sign}\left(\eta(X) - 1/2\right)$, so in particular, when $\mathrm{sign}(f(X)) \neq \mathrm{sign}\left(\eta(X) - 1/2\right)$, we have $H^-(\eta(X)) \leq \mathbb{E}[\phi(Yf(X))|X]$.

Also whether the sign condition is satisfied or not,

$$\mathbb{E}\left[\phi(Yf(X))|X\right] \geq H(\eta(X)).$$

Thus, considering either value of the indicator shows that

$$\begin{aligned}\psi\left(R(f) - R^*\right) &\leq \mathbb{E}\left[\phi(Yf(X)) - H(\eta(X))\right] \\ &= R_\phi(f) - R_\phi^*.\end{aligned}$$

Classification calibration for convex ϕ

Extensions:

- Every classification-calibrated ϕ is an upper bound on loss: there is a c such that $c\phi(\alpha) \geq 1[\alpha \leq 0]$.
- Flatter ϕ (smaller Bregman divergence at 0) gives a tighter bound on $R(f) - R^*$ in terms of $R_\phi(f) - R_\phi^*$.
- Under a low noise condition (that is, $\eta(X)$ is unlikely to be near $1/2$), the bound on excess risk in terms of excess ϕ-risk is improved.