Statistical Approaches to Learning and Discovery

Week 4: Decision Theory and Risk Minimization

February 3, 2003

Recall From Last Time

Bayesian expected loss is

$$\rho(\pi, a) = E_\pi[L(\theta, a)] = \int L(\theta, a)\, dF^\pi(\theta)$$

Conditioned on evidence in data X, we average with respect to the posterior:

$$\rho(\pi, a \,|\, X) = E_{\pi(\cdot\,|\,X)}[L(\theta, a)] = \int L(\theta, a)\, p(\theta \,|\, X)$$

Classical formulation: $\delta : \mathcal{X} \longrightarrow \mathcal{A}$ a decision rule, *risk function*

$$R(\theta, \delta) = E_X[L(\theta, \delta(X)] = \int_{\mathcal{X}} L(\theta, \delta(X))\, dF^X(x)$$

Bayes Risk

For a prior π, the *Bayes risk* of a decision function is

$$r(\pi, \delta) = E_\pi[R(\theta, \delta)] = E_\pi\left[E_X[L(\theta, \delta(X))]\right]$$

Therefore, the classical and Bayesian approaches define different risks, by averaging:

- Bayesian expected loss: Averages over θ
- Risk function: Averages over X
- Bayes risk: Averages over both X and θ

Example

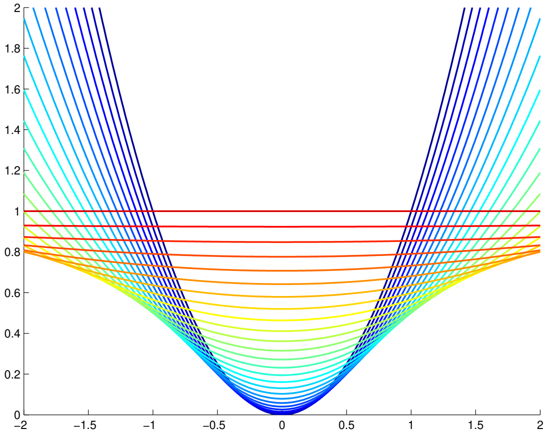
Take $X \sim \mathcal{N}(\theta, 1)$, and problem of estimating θ under square loss $L(\theta, a) = (a - \theta)^2$. Consider decision rules of the form $\delta_c(x) = cx$.

A calculation gives that

$$R(\theta, \delta_c) \;= c^2 + (1 - c)^2\theta^2$$

Then δ_c is inadmissible for $c > 1$, and admissible for $0 \leq c \leq 1$.

Example (cont.)



Risk $R(\theta, \delta_c)$ for admissible decision functions $\delta_c(x) = cx$, $c \leq 1$, as a function of θ. The color corresponds the associated minimum Bayes risk.

Example (cont.)

Consider now $\pi = \mathcal{N}(0, \tau^2)$. Then the Bayes risk is

$$r(\pi, \delta_c) = c^2 + (1-c)^2 \tau^2$$

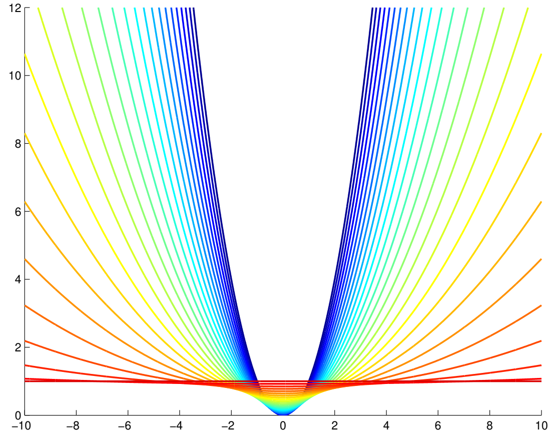
Thus, the best Bayes risk is obtained by the Bayes estimator δ_{c^*} with

$$c^* = \frac{\tau^2}{1+\tau^2}$$

(This is the same value of the Bayes risk of π.) That is, each δ_c is Bayes for the $\mathcal{N}(0, \tau_c^2)$ prior with

$$\tau_c = \sqrt{\frac{c}{1-c}}$$

Example (cont.)



At a larger scale, it becomes clearer that the decision function with $c = 1$ is minimax. It corresponds to the (improper) prior $\mathcal{N}(0, \tau^2)$ with $\tau \to \infty$.

Bayes Actions

$\delta^\pi(x)$ is a *posterior Bayes action for x* if it minimizes

$$\int_\Theta L(\theta, a)\, p(\theta \mid x)\, d\theta$$

Equivalently, it minimizes

$$\int_\Theta L(\theta, a)\, p(x \mid \theta)\, \pi(\theta)\, d\theta$$

Need not be unique.

Equivalence of Bayes actions and Bayes decision rules

A decision rule δ^π minimizing the Bayes risk $r(\pi, \delta)$ can be found "pointwise," by minimizing

$$\int_\Theta L(\theta, a)\, p(x \,|\, \theta)\, \pi(\theta)\, d\theta$$

for each x. So, the two problems are equivalent.

Classes of Loss Function

Three distinguished classes of problems/loss functions

- Regression: squared loss and relatives
- Classification: zero-one loss
- Density estimation: log-loss

Special Case: Squared Loss

For $L(\theta, a) = (\theta - a)^2$, the Bayes decision rule is the posterior mean

$$\delta^\pi(x) = E[\theta \,|\, x]$$

For weighted squared loss, $L(\theta, a) = w(\theta)(\theta - a)^2$, the Bayes decision rule is weighted posterior mean (when a is unrestricted):

$$\delta^\pi(x) = \frac{\int_\Theta \theta\, w(\theta)\, f(x \,|\, \theta)\, \pi(\theta)\, d\theta}{\int_\Theta w(\theta)\, f(x \,|\, \theta)\, \pi(\theta)\, d\theta}$$

Note: w acts like a prior here

We will see later how L^2 case—posterior mean—applies to some classification problems, in particular learning with labeled/unlabeled

data.

Special Case: Linear Loss

For $L(\theta, a) = |\theta - a|$, the Bayes decision rule is a posterior median.

More generally, for

$$L(\theta, a) = \begin{cases} c_0(\theta - a) & \theta - a \geq 0 \\ c_1(a - \theta) & \theta - a < 0 \end{cases}$$

a $\frac{c_0}{c_0+c_1}$-fractile of posterior $p(\theta \,|\, x)$ is a Bayes estimate.

Generic Learning Problem

In machine learning we're often interested in prediction.

Given input $X \in \mathcal{X}$, what is output $Y \in \mathcal{Y}$

Incur loss $L(X, Y, f(X))$ due to predicting $f(X)$ when input is X and true output is Y.

Generic Learning Problem (cont.)

Given a training set $(X_1, Y_1), \ldots (X_n, Y_n)$ and possibly unlabeled data $(X_1', \ldots X_m')$ determine $f : \mathcal{X} \longrightarrow \mathcal{Y}$ from some family $\mathcal{F}$.

Thus, a learning algorithm is a mapping

$$\mathcal{A} : \bigcup_{n \geq 0} (\mathcal{X} \times \mathcal{Y})^n \longrightarrow \mathcal{F}$$

in the supervised case and

$$\mathcal{A} : \bigcup_{n \geq 0, m \geq 0} (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X}^m \longrightarrow \mathcal{F}$$

in the semi-supervised case of labeled/unlabeled data.

Criterion for Success

Average loss on new data

$$\begin{aligned} R[f] &= E[L(X, Y, f(X))] \\ &= \int_{\mathcal{X} \times \mathcal{Y}} L(x, y, f(x))\, dP(x, y) \end{aligned}$$

for some (unknown) measure P on $\mathcal{X} \times \mathcal{Y}$.

Generalization error of a learning algortithm $\mathcal{A}$ is

$$R[\mathcal{A}] - \inf_{f \in \mathcal{F}} R[f]$$

Note: not assuming correctness of the model–risk may be greater than the Bayes error rate.

Empirical Risk

Since P is not typically known, can't compute the risk, and work instead with the *empirical risk*

$$R_{\text{emp}}[f] = R_{\text{emp}}[f, (x^n, y^n)] = \frac{1}{n}\sum_{i=1}^{n} L(x_i, y_i, f(x_i))$$

How might this be modified to take into account unlabeled data $(X'_1, \ldots, X'_m)$?

Standard Example

Consider $\mathcal{F} = \{f(x) = \langle x, w, \rangle\}$, the set of linear functions, and squared error $L(x, y, f(x)) = (y - f(x))^2$.

Minimizing empirical risk:

$$R_{\text{emp}}[f] = \min_w \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, x_i\rangle)^2$$

Solution $w = (X^\top X)^{-1} X^\top y$.

Note: can use a set of basis functions $\phi_i(x)$.

Other Loss Functions (cont.)

Most natural loss function for classification is 0-1 loss:

$$L(x, y, f(x)) \quad = \quad \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise} \end{cases}$$

Can specify other off-diagonal costs, e.g. distance function $d(y, f(x))$.

Other Loss Functions (cont.)

L^2 loss strongly affected by outliers. L^1 loss more "forgiving," though not differentiable. Again assume a linear model

$$\min_w R_{\text{emp}}[f] = \min_w \frac{1}{n}\sum_{i=1}^{n} |y_i - \langle w, x_i\rangle|$$

Can transform to a linear program:

$$\begin{array}{rl} \text{minimize} & \displaystyle\frac{1}{n}\sum_{i=1}^{n}(z_i + z_i^*) \\ \text{subject to} & y_i - \langle x_i, w\rangle \le z_i \\ & y_i - \langle x_i, w\rangle \ge -z_i^* \end{array}$$

Other Loss Functions (cont.)

Median property: at an optimal solution, an equal number of points will have $y_i - f(x_i) > 0$ and $y_i - f(x_i) < 0$.

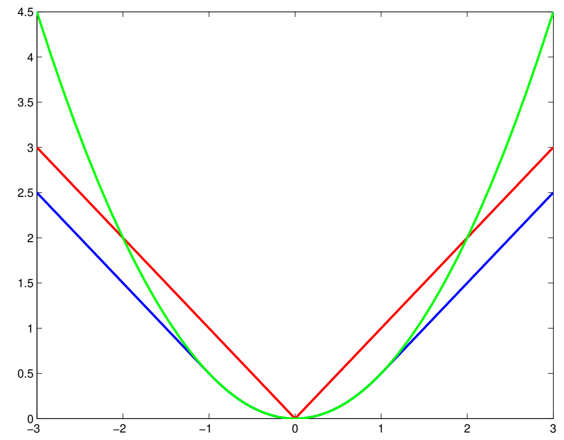
Other Loss Functions (cont.)

Huber's "robust loss" combines L^2 and L^1 for large errors

$$L_\sigma(x, y, f(x)) \;=\; \begin{cases} \frac{1}{2\sigma}(y - f(x))^2 & \text{if } |y - f(x)| \le \sigma \\ |y - f(x)| - \frac{\sigma}{2} & \text{otherwise} \end{cases}$$

(As we'll see, regularization is generally to be preferred over changing the loss function...but many learning algorithms do both)

Comparison of Loss Functions



Comparison of L^2 (green), L^1 (red) and Huber's robust loss (blue)

Probabilistic Error Models

Suppose we assume $Y = f(X) + \xi$ where $\xi \sim p_\theta$.

Then conditional probability $p(Y \mid f, X)$ can be computed in terms of $p_\theta(Y - f(X))$.

Note: error distribution could depend on X

Probabilistic Error Models (cont.)

Assume log-loss for the errors:

$$L(x, y, f(x)) = -\log p_\theta(y - f(x))$$

Then under the iid assumption, we have that

$$\begin{aligned} R_{\text{emp}}[f] &= \frac{1}{n}\sum_i L(x_i, y_i, f(x_i)) \\ &= -\frac{1}{n}\sum_i \log p_\theta(y_i - f(x_i)) + \text{constant} \end{aligned}$$

Looked at differently, the conditional distribution of y is given by

$$p(y \,|\, x, f) \quad \propto \quad \exp\left(-L(x, y, f(x))\right)$$

Consistency

For a *consistent* learning algorithm, we require that

$$\lim_{n\to\infty} P\left(R\left[\mathcal{A}(X^n, Y^n)\right] - R[f^*] > \epsilon\right) = 0$$

where the probability is w.r.t. the choice of training sample and $f^* = \arg\min_{f\in\mathcal{F}} R[f]$ achieves the minimum risk.

Relying on R_{emp} alone may require very large sample size to achieve small generalization error.

May also lead to ill-posed problems (non-unique, poorly conditioned). A small change in training data can lead to classifiers with very different expected risks (high variance)

Consistency (cont.)

Empirical risk $R_{\text{emp}}[f]$ converges to $R[f]$ for any *fixed* function f (e.g., McDiarmid's inequality)

But minimizing empirical risk gives *different* function for each sample. Showing consistency requires uniform convergence arguments.

As we'll discuss, such results and rates of convergence involve measures of complexity such as VC dimension or covering numbers (or some more recent notions...)

Regularization

Idea: want to restrict $f \in \mathcal{F}$ to some compact subset, e.g., $\Omega[f] \leq c$. May lead to difficult optimization problem.

Regularized risk is given by

$$A_{\Omega,\lambda}(X^n, Y^n) \quad = \quad \arg\min_{f\in F} \left(R_{\text{emp}}[f] + \lambda\Omega[f]\right)$$

where $\Omega : \mathcal{F} \longrightarrow \mathbb{R}$ is some (typically convex) function.

Then for appropriate choice of $\lambda \to 0$, regularization will achieve optimal risk $R[f^*]$ as $n \to \infty$

Bayesian Connection

If we assume a prior distribution on classifiers given by

$$\pi(f) \propto \exp\left(-n\lambda\Omega[f]\right)$$

then the posterior is given by

$$\begin{aligned} P(f \,|\, (X^n, Y^n)) &\propto \exp\left(-\sum_{i=1}^{n} L(X_i, Y_i, f(X_i))\right) \exp\left(-n\lambda\Omega[f]\right) \\ &= \exp\left(-R_{\text{emp}}[f] - \lambda\Omega[f]\right) \end{aligned}$$

so regularization corresponds to MAP estimation

We'll return to this when we discuss generative vs. discriminative models for learning, model selection