Decision Theory

Paolo Favaro
Universität Bern, Switzerland
favaro@inf.unibe.ch

1 Notation

Let X be a random variable with probability density distribution p_X and represent our *model* variables (for example, the class in a categorization problem or the coefficients of a polynomial in a data fitting problem). Also, let y be an *observation/measurement* of some function of X. For example, we could define a random variable Y such that $Y = f(X) + N$, where N represents zero-mean noise (independent of X), and y would be an instance of Y. Another example is, f(X,Y) = 0, where the expression is in implicit form. As a practical application, let us define $X = \{\theta_0, \ldots, \theta_n\}$, where θ_i is the coefficient corresponding to the monomial of order i in a polynomial expression. Let us also define $Y \in \mathbf{R}^2$. The function f in the implicit form can then be a polynomial, $f(X, Y) = Y_1 - \sum_{i=0}^{n} \theta_i Y_2^i$. Instances y of Y define measured samples (2D points) on the polynomial curve. An instance x of X defines a polynomial curve and the function f defines the relationship between the samples and the models.

2 Expected Loss Minimization

Given observations $y_1, \ldots, y_m$ (our data) one might be interested in computing an estimate $\hat{x}$ of the x instance (our model) that best describes them. The estimate $\hat{x}$ is also called a *decision rule* and maps an instance y to a model $\hat{x}(y)$. We assume that the relationship f between models X and data Y is known.

Towards this goal, let us define a *loss function* L between an instance x of X and its estimate $\hat{x}$ given y. The loss function L is typically small when the estimate $\hat{x}$ is close to x and large when different. For example, one could define $L(x, \hat{x}(y)) = |x - \hat{x}(y)|^2$. Thus, if we seek for the decision rule $\hat{x}$ that yields the smallest loss function L, we will tend to approach the true underlying model x. Then,

we can introduce the *Bayes risk* as the expectation

$$\begin{aligned} E_{X,Y}[L(X,\hat{x}(Y))] &= \int L(x,\hat{x}(y))p(x,y)dxdy \\ &= \int L(x,\hat{x}(y))p(y|x)p(x)dxdy \\ &= \int L(x,\hat{x}(y))p(x|y)p(y)dxdy \\ &= E_Y[E_{X|Y}[L(X,\hat{x}(Y))]] \end{aligned} \tag{1}$$

and we define the *posterior expected loss* as

$$E_{X|Y}[L(X,\hat{x}(Y))] = \int L(x,\hat{x}(y))p(x|y)dx \tag{2}$$

$$= \int L(x,\hat{x}(y))p(y|x)p(x)/p(y)dx. \tag{3}$$

Given y, we define the estimator $\tilde{x}$ as the decision rule that minimizes the Bayes risk

$$\tilde{x} = \arg\min_{\hat{x}} E_Y[E_{X|Y}[L(X,\hat{x}(Y))]]. \tag{4}$$

This is equivalent to defining the mapping element-wise via the posterior expected loss as

$$\tilde{x}(y) = \arg\min_{\hat{x}(y)} E_{X|Y}[L(X,\hat{x}(Y))]. \tag{5}$$

To keep the notation simple, from now on we will avoid showing the dependence of $\hat{x}$ with respect to y.

2.1 Example: Quadratic Loss

Let us compute the optimal estimator when $L(x,\hat{x}) = |x-\hat{x}|^2$. Then, the Bayes Risk is

$$\tilde{x} = \arg\min_{\hat{x}} \int |x-\hat{x}|^2 p(x,y)dxdy. \tag{6}$$

By using Calculus of Variations, we impose the first order condition for a minimum. That is, we set the gradient of the Bayes Risk with respect to $\hat{x}$ to 0, *i.e.*,

$$\frac{\delta E_{X,Y}[L(X,\hat{x})]}{\delta \hat{x}} = 2\int (\tilde{x}-x)p(x,y)dx = 0 \qquad \forall y. \tag{7}$$

By rearranging the previous equation we obtain

$$\tilde{x} = \frac{\int xp(x,y)dx}{\int p(x,y)dx} = \frac{E_{X,Y}[X]}{p(y)} = E_{X|Y}[X]. \tag{8}$$

Notice that the Bayes Risk at the optimal decision rule becomes

$$E_Y[E_{X|Y}[\left|X - E_{X|Y}[X]\right|^2]] = E_Y[\text{var}(X|Y)] \tag{9}$$

where $\text{var}(X|Y)$ is the variance of X given Y.

2.2 Example: Minkowski's Loss

A generalization of the quadratic loss function is the *Minkowski* loss

$$L_q = |x - \hat{x}|^q \tag{10}$$

with $q \geq 0$. When $q = 2$ then we have the quadratic loss presented in the previous section. If $q = 1$ then we have the conditional median. In fact, the minimization problem becomes

$$\tilde{x} = \arg\min_{\hat{x}} \int |x - \hat{x}| p(x,y)dxdy \tag{11}$$

and by using Calculus of Variations we get

$$\frac{\delta E_{X,Y}[L_1(X,\hat{x})]}{\delta \hat{x}} = \left(\int_{x|x \succ \hat{x}} xp(x|y)dx - \int_{x|x \prec \hat{x}} xp(x|y)dx \right) p(y) = 0 \qquad \forall y. \tag{12}$$

This means that $\hat{x}$ must satisfy

$$\int_{x|x \succ \hat{x}} xp(x|y)dx = \int_{x|x \prec \hat{x}} xp(x|y)dx \tag{13}$$

which is, by definition, the *conditional median*. Another important case is when $q \mapsto 0$. In this case the loss L_q tends to the constant 1 everywhere except in $\hat{x} \equiv x$ where it is 0. The optimal solution is then the decision rule corresponding to the maximum of $p(X|Y)$, *i.e.*,

$$\tilde{x} = \arg\max_{\hat{x}} p(\hat{x}|y) \tag{14}$$

which is the so-called *Maximum a Posteriori* (MAP) estimate. The same problem formulation can be obtained by considering the loss function $L(x, \hat{x}) = 1 - \delta(x - \hat{x})$. In this case the optimal estimator becomes

$$\tilde{x} = \arg\min_{\hat{x}} 1 - p(\hat{x}|y) = \arg\max_{\hat{x}} \frac{p(y|\hat{x})p(\hat{x})}{p(y)} = \arg\max_{\hat{x}} p(y|\hat{x})p(\hat{x}). \tag{15}$$

2.3 Generative and Discriminative Approaches

The calculation of the optimal estimator $\tilde{x}$ requires the minimization of the posterior expected loss. This boils down to the calculation of an integral which could be done by using either eq. (2) or eq. (3). When using the first expression, we need to provide an expression for the posterior $p(x|y)$. This approach is called *discriminative* as it focuses on the direct discrimination between the models.

The second expression requires an expression for the posterior $p(y|x)$ and $p(x)$. In this case the approach is called *generative* as it relies on knowledge of how the observations y are generated by the model defined by x. The probability $p(y)$ can then be obtained by using marginalization, *i.e.*, $p(y) = \int p(y|x)p(x)dx$. The probabilities then can be combined thanks to Bayes rule to yield

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{16}$$

Discriminative approaches do not require knowledge of the underlying relation f between data and models, thus, they cannot be used to generate samples of y given x. However, this can also be an advantage. Since the same posterior $p(x|y)$ can be generated by different relations between data and models, a discriminative approach makes fewer assumptions on the estimation task and might be more robust as a result.

Generative approaches typically do better when the relationship between data and models is known as the data fitting can be tighter than with discriminative approaches. Also, these approaches tend to be more easily adapted to unsupervised learning problems.

2.4 Variational Bayes

In the Variational Bayes approach we look for an explicit representation of $p(X|Y)$. We start from the data Y and define a function q such that $q(X) \geq 0$ and $\int q(X)dX = 1$. Then we have

$$\begin{aligned}
\log p(Y) &= \log p(Y) \int q(X)dX \\
&= \int q(X) \log p(Y)dX \\
&= \int q(X) \log \frac{p(X,Y)}{p(X|Y)}dX \\
&= \int q(X) \log \frac{p(X,Y)q(X)}{p(X|Y)q(X)}dX \\
&= \int q(X) \log \frac{p(X,Y)}{q(X)}dX - \int q(X) \log \frac{p(X|Y)}{q(X)}dX \\
&= \mathcal{L}(q) + D(q|p)
\end{aligned} \tag{17}$$

where $D(q|p) = -\int q(X) \log \frac{p(X|Y)}{q(X)}dX$ is the Kullback-Leibler distance between $q(X)$ and $p(X|Y)$. Thus $D(q|p) \geq 0$ by definition. Since $\log p(Y)$ is fixed with respect to X, and therefore also with respect to $q(X)$, increasing $\mathcal{L}$ with respect to q will have to be compensated by an equal decrease of $D(q|p)$, and vice versa. Thus, the maximum of $\mathcal{L}$ is obtained as the minimum of $D(q|p)$. Because the minimum of the Kullback-Leibler distance $D(q|p)$ is achieved when its two arguments are identical, we have $q(X) = p(X|Y)$, which achieves our initial objective.

The maximization of $\mathcal{L}$ can be performed via an iterative method and also the calculation of the cost $\mathcal{L}$ can be made feasible by using simple forms of $q(X)$. One example is the *mean field* approximation where q is written in factorized form as $q(X) = \Pi_i q_i(X_i)$, where each $q_i \geq 0$ and $\int q_i(X_i)dX_i = 1$. If q is interpreted as a probability distribution, this approximation would be equivalent to assuming independence between the components of X.

Once an approximation $q(X)$ of $p(X|Y)$ has been computed, it is possible to evaluate the integral in eq. (5) explicitly

$$\begin{aligned}
\tilde{x} &= \arg\min_{\hat{x}} \int p(x|y)L(x,\hat{x})dx \\
&= \arg\min_{\hat{x}} \int q(x)L(x,\hat{x})dx.
\end{aligned} \tag{18}$$

If the above mean field approximation is used, we obtain

$$\tilde{x} = \arg\min_{\hat{x}} \int L(x,\hat{x})\Pi_i q_i(x_i)dx_i. \tag{19}$$

If the loss function factorizes as $L(x, \hat{x}) = \sum_i L_i(x_i, \hat{x}_i)$, then the problem simplifies to

$$\tilde{x} = \arg\min_{\hat{x}} \sum_i \int q_i(x_i) L_i(x_i, \hat{x}_i) dx_i \tag{20}$$

or, equivalently,

$$\tilde{x}_i = \arg\min_{\hat{x}_i} \int q_i(x_i) L_i(x_i, \hat{x}_i) dx_i. \tag{21}$$

The mean field approximation can then lead to a much more computationally feasible calculation.