Lecture 6

11. Decision theory

Recall from Section 8 that for a decision rule δ and observation $X = x$ we have (in Bayesian setting) the posterior risk

$$r(\delta \mid x) = \int_\Omega L(\theta, \delta(x))\mu_{\Theta|X}(d\theta \mid x),$$

where $L(\theta, \delta(x)) = \int_\aleph L(\theta, a)\delta(da; x)$ if δ is a randomized rule. If δ_0 is a decision rule such that for all x, $r(\delta_0 \mid x) < \infty$ and for all x and all decision rules δ $r(\delta_0 \mid x) \leq r(\delta \mid x)$, then δ_0 is called a *formal Bayes rule*.

There is also a weaker concept than a formal Bayes rule. Denote by μ_Θ the prior distribution of Θ. Together with $f_{X|\Theta}$ this specifies the predictive (marginal) distribution of X, μ_X. We call the function

$$r(\mu_\Theta, \delta) = \int_{\mathcal{X}} r(\delta \mid x)\mu_X(dx)$$

the *Bayes risk* and each δ that minimizes the Bayes risk is called a *Bayes rule* with respect to μ_Θ, assuming $r(\eta, \delta) < \infty$. The Bayes risk is the mean of the posterior risk, before observing $X = x$.

11.1. **Classical decision theory.**

In classical decision theory we condition on $\Theta = \theta$ and introduce the *risk function*

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x))\mu_{X|\Theta}(dx \mid \theta).$$

That is, the conditional mean of the loss, given $\Theta = \theta$. Here we would like to find a rule δ that minimizes the risk function simultaneously for all values of θ. As we saw in the last lecture there may not be a rule that minimizes the risk function simultaneously for all θ. Therefore we introduce the notion of admissible rules.

Definition 16. Let δ be a decision rule. If there exists a decision rule δ_1 such that $R(\theta, \delta_1) \leq R(\theta, \delta)$ for all θ with strict inequality for some θ, then we say δ is in-admissible and it is dominated by δ_1. Otherwise, δ is admissible.

Of course, one should not use in-admissible decision rules.

As a weaker criterion one can, as in the Bayesian setting, take a prior distribution μ_Θ for Θ and try to minimize

$$\int_\Omega R(\theta, \delta)\mu_\Theta(d\theta).$$

Note that by Fubini's theorem we have

$$\begin{aligned}
\int_\Omega R(\theta, \delta)\mu_\Theta(d\theta) &= \int_\Omega \int_{\mathcal{X}} L(\theta, \delta(x))\mu_{X|\Theta}(dx \mid \theta)\mu_\Theta(d\theta) \\
&= \int_{\mathcal{X}} \int_\Omega L(\theta, \delta(x))\mu_{\Theta|X}(d\theta \mid x)\mu_X(dx) \\
&= \int_{\mathcal{X}} r(\delta \mid x)\mu_X(dx) = r(\eta, \delta)
\end{aligned}$$

which is the Bayes risk with respect to μ_Θ.

Minimax rules. For a given problem there might be many admissible decision rules, but we may not be able to find one which dominates all the others. In that case we need a criteria to decide which rule to take. We have already seen the possibility of choosing a Bayes rule with respect a some prior distribution η. A different criteria is the following.

Definition 17. A decision rule δ_0 is called *minimax* if

$$\sup_{\theta\in\Omega} R(\theta,\delta_0) \le \inf_{\delta}\sup_{\theta\in\Omega} R(\theta,\delta).$$

That is, a minimax has the smallest upper bound of the risk function. That is, we prepare for the worst possible θ and choose the rule which has the smallest risk for this worst θ. One could ask how minimax rules are connected to Bayes rules. If λ is a prior for Θ we have

$$r(\lambda,\delta) = \int_{\Omega} R(\theta,\delta)\lambda(d\theta).$$

Hence, if λ puts all its mass on those θ that maximizes $R(\theta,\delta)$ we see that

$$\sup_{\lambda} r(\lambda,\delta) = \sup_{\theta} R(\theta,\delta).$$

This choice of λ depends on the decision rule δ.

Definition 18. A prior distribution λ_0 for Θ is *least favorable* if $\inf_\delta r(\lambda_0,\delta) = \sup_\lambda \inf_\delta r(\lambda,\delta)$.

That is, λ_0 is a prior such that the corresponding Bayes rule has the highest possible risk.

For any fixed prior λ_0 and decision rule δ_0 we have

$$\inf_{\delta} r(\lambda_0,\delta) \le r(\lambda_0,\delta_0) \le \sup_{\lambda} r(\lambda,\delta_0).$$

Therefore we can introduce the following concept.

Definition 19. Let

$$V_- \equiv \sup_{\lambda}\inf_{\delta} r(\lambda,\delta) \le \inf_{\delta}\sup_{\lambda} r(\lambda,\delta) = \inf_{\delta}\sup_{\theta} R(\theta,\delta) \equiv V^-.$$

Then V_- is the *maximin* value of the decision problem and V^- is the *minimax* value of the decision problem.

How can we check that a rule is minimax and a prior least favorable?

Theorem 16. *If δ_0 is a Bayes rule with respect to λ_0 and $R(\theta,\delta_0) \le r(\lambda_0,\delta_0)$ for all θ, then δ_0 is minimax and λ_0 is least favorable.*

Proof. Since

$$V^- \le \sup_{\theta} R(\theta,\delta_0) \le r(\lambda_0,\delta_0) = \inf_{\delta} r(\lambda_0,\delta) \le V_-$$

and $V_- \le V^-$ it must be that $V_- = V^-$ and the claim follows. □

The theorem gives you a condition to check but when can we actually find minimax rules. We will consider the case where Ω is finite, $\Omega = \{\theta_1,\dots,\theta_k\}$. In that case the risk function $R(\theta,\delta)$ for a given decision rule δ is just a vector in $\mathbb{R}^k$.

Definition 20. Suppose $\Omega = \{\theta_1, \dots, \theta_k\}$, let

$$R = \{z \in \mathbb{R}^k : z_i = R(\theta_i, \delta), i = 1, \dots, k, \text{ for some decision rule } \delta\}.$$

The set R is called the *risk set.* For any $C \subset \mathbb{R}^k$ the *lower boundary* is the set

$$\{z \in C^- : x_i \le z_i, i = 1 \dots, k \text{ and } x_i < z_i \text{ for some } i \text{ implies } x \notin C^-\}.$$

The lower boundary of the risk set is denoted ∂L. The risk set is closed from below if $\partial L \subset R$.

Lemma 3. *The risk set is convex.*

Proof. For $i = 1, 2$ let $z_i \in R$ be points that correspond to the decision rules δ_i and take $\lambda \in [0, 1]$. Then $\lambda z_1 + (1 - \lambda) z_2$ is the risk function of the randomized decision rule corresponding to taking δ_1 with probability λ and δ_2 with probability $1 - \lambda$. Hence, it belongs to the risk set R. □

Consider Example 3.72, p. 170 in Schervish "Theory of statistics".

Theorem 17 (Minimax theorem)**.** *Suppose the loss function is bounded from below and* Ω *is finite. Then* $\sup_\lambda \inf_\delta r(\lambda, \delta) = \inf_\delta \sup_\theta R(\theta, \delta)$ *and a least favorable prior* λ_0 *exists. If* R *is closed from below, then there exists a minimax rule that is a Bayes rule with respect to* λ_0*.*

Proof. For any real number s let $A_s = \{z \in \mathbb{R}^k : z_i \le s,\ i = 1, \dots, k\}$. That is, A_s is an orthant. It is closed and convex for each s. Take $s_0 = \inf\{s : A_s \cap R \neq \emptyset\}$. Then

$$s_0 = \inf_\delta \sup_\theta R(\theta, \delta).$$

Indeed, for each $z \in A_s \cap R$ there is a decision rule δ such that $\sup_\theta R(\theta, \delta) = \max_i R(\theta_i, \delta) \le s$. Taking inf over s corresponds exactly to taking inf over δ. Next note that the interior of A_{s_0} is convex and does not intersect R. The separating hyperplane theorem says that there exists a vector v and a real number c such that $v^T z \ge c$ for each $z \in R$ and $v^T z \le c$ for each x in the interior of A_{s_0}. It is necessary that each coordinate of v satisfies $v_j \ge 0$. Otherwise, if $v_j < 0$ we can find a sequence x_n in the interior of A_{s_0} with $\lim_n x_{ni} = -\infty$ and all other $x_{nj} = s_0 - \varepsilon$ and then $\lim_n v^T x_n = \infty > c$, which is a contradiction. If we put $\lambda_{0j} = v_j / \sum_{j=1}^k v_j$ we get a probability measure on Ω which is least favorable. Indeed, since $(s_0, \dots, s_0)$ is in the closure of the interior of A_{s_0} it follows that $c \ge s_0 \sum_{j=1}^k v_j$ and we have

$$\inf_\delta r(\lambda_0, \delta) = \inf_{z \in R} \lambda_0^T z \ge \frac{c}{\sum_{j=1}^k v_j} \ge s_0 = \inf_\delta \sup_\theta R(\theta, \delta)$$

This shows that λ_0 is least favorable.

We were not able to cover the proof that there exists a minimax rule. We refer to the book (Schervish, p.173). □

11.2. **On finding a formal Bayes rule.** In Bayesian decision theory the following is a good way to find a deterministic formal Bayes rule.

(1) Take $x \in \mathcal{X}$.
(2) Find $a \in \aleph$ that minimizes $\int_\Omega L(\theta, a) \mu_{\Theta|X}(d\theta \mid x)$.
(3) Put $\delta(x) = a$.
(4) Repeat for all x.

However, it is not always that a formal Bayes rule exists, for instance the minimum in step (2) may not exist in $\aleph$. Here is an example

Example 23. Let $X \sim N(\theta, 1)$ and $\Theta \sim N(0,1)$ where $\Omega = \mathbb{R}$. Then the posterior is $N(x/2, 1/2)$. Let the action space be $\aleph = \mathbb{R}$ and the loss function $L(\theta, a) = 0$ if $a \geq \theta$, $L(\theta, a) = 1$ if $a < \theta$. That is, a loss occurs if our guess of θ is below θ. Then for any x

$$\int_\Omega L(\theta, a)\mu_{\Theta|X}(d\theta \mid x) = \mu_{\Theta|X}(\Theta > a \mid x) = 1 - \Phi\Big(\frac{a - x/2}{1/\sqrt{2}}\Big).$$

This converges to 0 as $a \to \infty$, so the risk is minimized at $a = \infty$ but this is not in the action space $\aleph$. For this example no formal Bayes rule exists.

12. The Neyman-Pearson fundamental lemma

Definition 21. A class $\mathcal{C}$ of decision rules is *complete* if for every $\delta \notin \mathcal{C}$ there exists $\delta_0 \in \mathcal{C}$ that dominates δ, i.e. $R(\theta, \delta_0) \leq R(\theta, \delta)$ $\forall \theta$ with strict inequality for some θ.

A class in *minimal complete* if no proper subclass is also complete.

To see the relation to admissible decision rules, we have the following:

Lemma 4. *A minimal complete class consists exactly of the admissible decision rules.*

Proof. First we show that δ admissible implies $\delta \in \mathcal{C}$. Indeed, if $\delta \notin \mathcal{C}$ then there exists $\delta_0 \in \mathcal{C}$ that dominates δ which contradicts that δ is admissible.

For the other inclusion we need to show that $\delta \in \mathcal{C}$ implies δ is admissible. Suppose it is not admissible. Then exists a dominating rule δ_1. Either $\delta_1 \in \mathcal{C}$ or $\delta_1 \notin \mathcal{C}$. In the first case put $\delta_2 = \delta_1$. In the second, there is $\delta_2 \in \mathcal{C}$ that dominates δ_1. Thus, in both cases $\delta_2 \in \mathcal{C}$ dominates δ. If δ' is a rule that is dominated by δ, then it is also dominated by δ_2. This implies that $\mathcal{C} \setminus \{\delta\}$ is complete. This is a contradiction because we assumed that $\mathcal{C}$ is minimal complete. Hence, δ is admissible. $\square$

There is one, simple case, where a minimal complete class can be found. This is called the Neyman-Pearson fundamental lemma.

Theorem 18. *Let $\Omega = \aleph = \{0,1\}$, $L(0,0) = L(1,1) = 0$, $L(1,0) = k_1 > 0$, and $L(0,1) = k_0 > 0$. Let $f_i(x) = dP_i/d\nu$ where ν is $P_0 + P_1$. For δ, a decision rule, let $\phi(x) = \delta(\{1\}; x)$ be the test function of δ. Let $\mathcal{C}$ be the class of rules with test functions of the form below:*

For each $k \in (0, \infty)$ and each function $\gamma : \mathcal{X} \to [0,1]$,

$$\phi_{k,\gamma}(x) = \begin{cases} 1, & f_1(x) > kf_0(x), \\ \gamma(x), & f_1(x) = kf_0(x), \\ 0, & f_1(x) < kf_0(x). \end{cases}$$

For $k = 0$,

$$\phi_0(x) = \begin{cases} 1, & f_1(x) > 0, \\ 0, & f_1(x) = 0. \end{cases}$$

For $k = \infty$,

$$\phi_\infty(x) = \begin{cases} 1, & f_0(x) = 0, \\ 0, & f_0(x) > 0. \end{cases}$$

Then $\mathcal{C}$ is a minimal complete class.

Before we prove the result let us see what the decision rules are. The decision rules are asssociated with a threshold $k \in [0, \infty]$.

- To $k = 0$ there corresponds one decision rule which says "choose $a = 1$ if $f_1(x) > 0$ and $a = 0$ otherwise".
- To $k = \infty$ there corresponds one decision rule which says "choose $a = 1$ if $f_0(x) = 0$ and $a = 0$ otherwise".
- To each $k \in (0, \infty)$ there are lots of decision rules. They all say that $a = 1$ should be chosen if it is sufficiently likely that $\theta = 1$. That is: "choose $a = 1$ if $f_1(x) > kf_0(x)$, choose $a = 0$ if $f_1(x) < kf_0(x)$, and in the event that we cannot decide $f_1(x) = kf_0(x)$ we choose $a = 1$ with probability $\gamma(x)$ where γ is some function $\gamma : \mathcal{X} \to [0, 1]$".

Example 24. The Neyman-Pearson lemma can be used when deciding between competing models. Suppose we have two competing models for the distribution of X given by continuous densities f_0 and f_1 w.r.t. Lebesgue measure. Based on observing $X = x$ we have to decide which is the more appropriate one. Decisions are $a = 1$ "f_1 is correct density" and $a = 0$ "f_0 is correct". The Neyman-Pearson lemma says that the admissible rules (the minimal complete class) are of the form: for $k \in (0, \infty)$ choose $a = 1$ if $f_1(x) > kf_1(x)$ and $a = 0$ if $f_1(x) < kf_0(x)$. There is no need to specify the case $f_1(x) = kf_0(x)$ since this even has probability zero. Also the cases $k = 0$ or ∞ corresponds to "always choose $a = 1$" and "always choose $a = 0$". None of these seem very desirable.

Example 25. If we continue the above example when $f_0(x) = \lambda_0^{-1} e^{-\lambda_0 x}$ and $f_1(x) = \lambda_1^{-1} e^{-\lambda_1 x}$ we see that we choose $a = 1$ if

$$\frac{f_1(x)}{f_0(x)} > k \iff x \le \frac{\log \lambda_1 - \log \lambda_0 + \log k}{\lambda_1 - \lambda_0}.$$

You can think of the case $k = 1$ as the fair case where we choose the model which is most likely. $k > 1$ penalizes choosing $a = 1$ whereas $k < 1$ penalizes choosing $a = 0$.

Proof of Neyman-Pearson's fundamental lemma. The proof is outlined as follows. First we consider a larger class $\mathcal{C}'$ which contains $\mathcal{C}$ and show that $\mathcal{C}'$ is complete. Then we will show that each rule in $\mathcal{C}'$ is dominated by a rule in $\mathcal{C}$ and that $\mathcal{C}$ is minimal complete.

The class $\mathcal{C}'$ consists of the class $\mathcal{C}$ and in addition the rules with testfunction of the form

$$\phi_{0,\gamma}(x) = \begin{cases} 1, & f_1(x) > 0, \\ \gamma(x), & f_1(x) = 0. \end{cases}$$

We will show that $\mathcal{C}'$ is complete. That is, for any rule $\delta \notin \mathcal{C}'$ there is a $\delta' \in \mathcal{C}'$ that dominates δ. Let $\delta \notin \mathcal{C}'$ be a rule with test function ϕ and put

$$\alpha = R(0, \delta) = \int_{\mathcal{X}} [L(0,0)(1 - \phi(x)) + L(0,1)\phi(x)] f_0(x)\nu(dx) = \int k_0 \phi(x) f_0(x) \nu(dx).$$

We will now try to find a rule $\delta' \in \mathcal{C}'$ with $R(0, \delta') = \alpha = R(0, \delta)$ and $R(1, \delta') < R(1, \delta)$. We define the function

$$g(k) = \int_{\{f_1(x) \ge kf_0(x)\}} k_0 f_0(x) \nu(dx).$$

Note that if $\gamma(x) = 1$ for all x and δ' has test function $\phi_{k,\gamma}$ then $g(k) = R(0, \delta')$. We claim that he function g has the following properties:

- $g(k) \to 0$ as $k \to \infty$.
- $g(0) = k_0 \geq \alpha$.
- $g(k)$ is continuous from the left and has limit from the right.

Note that $f_1(x) < \infty$ ν-a.e. and the set $\{f_1(x) \geq kf_0(x)\}$ decreases to $\emptyset$ with k. Hence $g(k) \to 0$ as $k \to \infty$. For the second claim,

$$g(0) = \int_{\mathcal{X}} k_0 f_0(x)\nu(dx) = k_0 \geq \alpha.$$

Let us show that g is left continuous. We have that

$$\bigcap_{k<m, k\in\mathbb{Q}} \{x : f_1(x) \geq kf_0(x)\} = \{x : f_1(x) \geq mf_0(x)\}.$$

The monotone convergence theorem gives

$$\lim_{k\uparrow m} g(k) = g(m),$$

We see that g is continuous from the left. To see is has limits from the right note

$$\bigcup_{k>m, k\in\mathbb{Q}} \{x : f_1(x) \geq kf_0(x)\} = \{x : f_1(x) > mf_0(x)\} \cup \{x : f_0(x) = 0\},$$

and since g is bounded the monotone convergence theorem implies

$$\lim_{k\downarrow m} g(k) = \int_{\{f_1(x) > mf_0(x)\}} k_0 f_0(x)\nu(dx)$$

so the limit from the right exists.

Note that if $\gamma(x) = 0$ for all x and δ' is a rule with test function $\phi_{m,\gamma}$, then $R(0, \delta') = \lim_{k\downarrow m} g(k)$. Since g is left continuous one of two cases can occur.

(i) either $g(k)$ decreases continuously to the level α, or
(ii) $g(k)$ jumps from a level above α to a level at most α.

In the first case there is a smallest k such that $g(k) = \alpha$ and we put $k^* = \inf\{k : g(k) = \alpha\}$. In the second case, there is a largest k such that $g(k) > \alpha$ and we put $k^* = \sup\{k : g(k) > \alpha\}$. In the case $\alpha = 0$ it is possible that $k^* = \infty$. If $\alpha > 0$ we must have $k^* < \infty$ because $g(k) \downarrow 0$ as $k \to \infty$. We will now construct a decision rule δ' with test function $\phi_{k^*,\gamma}$. There are three cases to consider:

(1) $\alpha = 0$ and $k^* < \infty$,
(2) $\alpha = 0$ and $k^* = \infty$,
(3) $\alpha > 0$ and $k^* < \infty$.

We proceed as follows. In each case 1, 2, and 3, we show that we can choose γ such that $R(0, \delta') = R(0, \delta) = \alpha$ and then that $R(1, \delta') < R(1, \delta)$.

Case 1: Take $\gamma(x) = 0$ for all x. Then

$$R(0, \delta') = \lim_{k\downarrow k^*} g(k) = \alpha = R(0, \delta).$$

Define

$$h(x) = [\phi_{k^*,\gamma}(x) - \phi(x)][f_1(x) - k^* f_0(x)].$$

We know that $\phi_{k^*,\gamma}(x) = 1 \geq \phi(x)$ on $\{x : f_1(x) - k^* f_0(x) > 0\}$ and $\phi_{k^*,\gamma}(x) = 0 \leq \phi(x)$ on $\{x : f_1(x) - k^* f_0(x) < 0\}$. Since ϕ is not of the form $\phi_{k,\gamma}$ for any k

and γ there must be a set B such that $\nu(B) > 0$ and $h(x) > 0$ on B. Using that $f_0(x) + f_1(x) = 1$ (since $\nu = P_0 + P_1$) we get

$$\begin{aligned}
0 &< \int_B h(x)\nu(dx) \le \int h(x)\nu(dx) \\
&= \int [\phi_{k^*,\gamma}(x) - \phi(x)] f_1(x)\nu(dx) - k^* \int [\phi_{k^*,\gamma}(x) - \phi(x)] f_0(x)\nu(dx) \\
&= \int [\phi_{k^*,\gamma}(x) - \phi(x)] f_1(x)\nu(dx) + \frac{k^*}{k_0}(\alpha - \alpha) \\
&= \frac{1}{k_1}[R(1,\delta) - R(1,\delta')].
\end{aligned}$$

Hence $R(1,\delta) < R(1,\delta')$.

Case 2: In this case

$$R(0,\delta') = \int k_0 \phi_\infty(x) f_0(x)\nu(dx) = 0 = \alpha.$$

Then since $0 = \alpha = R(0,\delta)$, $\phi(x) = 0$ for all x such that $f_0(x) > 0$. Then

$$\begin{aligned}
R(1,\delta) &= k_1 P_1(f_0(X) > 0) + k_1 \int_{\{x: f_0(x)=0\}} [1 - \phi(x)] f_1(x)\nu) dx) \\
&> k_1 P_1(f_0(X) > 0) = R(1,\delta').
\end{aligned}$$

Case 3: If $g(k^*) = \alpha$ we set $\gamma(x) = 1$ for all x, because then $R(0,\delta') = g(k^*) = \alpha$. If $g(k^*) > \alpha$ put

$$v = \lim_{k \downarrow k^*} g(k) \le \alpha.$$

In this case, g is discontinuous at k^* and

$$k_0 P_0(f_1(X) = k^* f_0(X)) = g(k^*) - v > \alpha - v \ge 0.$$

For x such that $f_1(x) = k^* f_0(x)$ we define

$$0 \le \gamma(x) = \frac{\alpha - v}{g(k^*) - v} < 1.$$

Then it follows that

$$\begin{aligned}
R(0,\delta') &= \int k_0 \phi_{k^*,\gamma}(x) f_0(x)\nu(dx) \\
&= v + \int_{\{x: f_1(x) = k^* f_0(x)\}} k_0 \frac{\alpha - v}{g(k^*) - v} f_0(x)\nu(dx) \\
&= v + \frac{\alpha - v}{g(k^*) - v} k_0 P_0(f_1(X) = k^* f_0(X)) = \alpha.
\end{aligned}$$

To see that $R(1,\delta') < R(1,\delta)$ we can proceed exactly as in Case 1 because k^* is finite. This finishes the proof that $\mathcal{C}'$ is complete.

To reduce from $\mathcal{C}'$ to $\mathcal{C}$ we need to show that if $\delta \in \mathcal{C}' \setminus \mathcal{C}$ then there is a rule $\delta' \in \mathcal{C}$ that dominates δ. This will show that $\mathcal{C}$ is a complete class.

Take $\delta' \in \mathcal{C}' \setminus \mathcal{C}$. Then the test function is $\phi_{0,\gamma}$ for some $\gamma : \mathcal{X} \to [0,1]$ such that $P_0(\gamma(X) > 0) > 0$. Let δ_0 be the test function with test function ϕ_0. Since $f_1(x) = 0$

for all x in the set $A = \{x : \phi_{0,\gamma}(x) \neq \phi_0(x)\}$ it follows that $R(1, \delta) = R(1, \delta_0)$. However,

$$\begin{aligned} R(0, \delta) &= k_0 E_0[\gamma(X) I_A(X)] + k_0 P_0(f_1(X) > 0) \\ &= k_0 E_0[\gamma(X) I_A(X)] + R(0, \delta_0) > R(0, \delta_0). \end{aligned}$$

Hence δ_0 dominates δ. It only remains to show that no element in $\mathcal{C}$ is dominated by any other element in $\mathcal{C}$. This shows the minimality of the class. The proof of this final step is an exercise (Problem 29, p. 212). □