T2.6 Teoría de la decisión Bayesiana

Índice

1 Conceptos básicos

Inferencia Bayesiana: cálculo de la **posterior** $p(H \mid \boldsymbol{x})$ mediante la regla de Bayes actualizar nuestras creencias sobre cantidades ocultas H a partir de datos $\boldsymbol{x}$

Teoría de la decisión Bayesiana: usa la inferencia para decidir cuál es la mejor de las posibles **acciones** a realizar

1.1 Conceptos básicos

Agente: debe escoger una acción de un conjunto de acciones posibles, $\mathcal{A}$

Estado de la naturaleza: $h \in \mathcal{H}$, condiciona los costes y beneficios que se derivan de tomar cada acción posible

Función de pérdida: indica el coste incurrido al tomar la acción $a \in \mathcal{A}$ cuando el estado de la naturaleza es $h \in \mathcal{H}$

$$\ell(h, a)$$

Riesgo (pérdida) esperado a posteriori: de a tras observar $\boldsymbol{x}$

$$R(a \mid \boldsymbol{x}) = \mathbb{E}_{p(h \mid \boldsymbol{x})}[\ell(h, a)] = \sum_{h \in \mathcal{H}} \ell(h, a)\, p(h \mid \boldsymbol{x})$$

Política óptima o estimador de Bayes: obtiene una acción de mínimo riesgo por cada observación posible

$$\pi^*(\boldsymbol{x}) = \operatorname*{argmin}_{a \in \mathcal{A}} \; R(a \mid \boldsymbol{x})$$

1.2 Conceptos (un poco menos) básicos

Función de utilidad: deseabilidad de cada acción posible en cada posible estado, esto es, riesgo con signo cambiado

$$U(h,a)=-\ell(h,a)$$

Principio de utilidad esperada máxima: estimador de Bayes expresado en términos de utilidad

$$\pi^*(\boldsymbol{x})=\operatorname*{argmax}_{a\in\mathcal{A}}\;\mathbb{E}_h[U(h,a)]$$

Sensibilidad al riesgo: asumimos que el agente es **neutral,** esto es, insensible al riesgo, pero podría no ser así; por ejemplo, nos da igual obtener 50 EUR con seguridad, o con 50% de probabilidad de 0 y 100 EUR

2 Problemas de clasificación

2.1 Problemas de clasificación: pérdida 01

Estados de la naturaleza y acciones: etiquetas de clase, $\mathcal{H} = \mathcal{Y} = \{1, \dots, C\} \qquad \text{y} \qquad \mathcal{A} = \mathcal{Y}$

Pérdida 01 para dos clases: $\mathcal{Y} = \{0, 1\}$

$$\ell_{01}(y^*, \hat{y}) = \left[\begin{array}{c|cc} & \hat{y}=0 & \hat{y}=1 \\ \hline y^*=0 & 0 & 1 \\ y^*=1 & 1 & 0 \end{array}\right] = \mathbb{I}(y^* \neq \hat{y})$$

Pérdida esperada a posteriori: la probabilidad de error a posteriori es uno menos la de acertar a posteriori

$$R(\hat{y} \mid \boldsymbol{x}) = \sum_y \ell_{01}(y, \hat{y})\, p(y \mid \boldsymbol{x}) = \sum_{y \neq \hat{y}} p(y \mid \boldsymbol{x}) = 1 - p(\hat{y} \mid \boldsymbol{x})$$

Esstimador de Bayes: **estimador máximo a posteriori (MAP),** esto es, la etiqueta más probable o **moda** de la probabilidad a posteriori

$$\pi(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \; p(y \mid \boldsymbol{x})$$

2.2 Matrices de confusión multiclase

Datos: conjunto de pares etiqueta real-predicha, $\mathcal{D} = \{(y_m, \hat{y}_m)\}_{m=1}^M$, obtenidos al clasificar M muestras (de test)

Matriz de confusión para C clases: $\mathbf{M} = [M_{y,\hat{y}}]$ con $M_{y,\hat{y}} = \sum_m \mathbb{I}(y_m = y)\mathbb{I}(\hat{y}_m = \hat{y})$

y	$\hat{1}$	$\hat{2}$	$\cdots$	$\hat{C}$	**Suma fila**
1	$M_{1,\hat{1}}$	$M_{1,\hat{2}}$	$\cdots$	$M_{1,\hat{C}}$	$M_{1,:}$
2	$M_{2,\hat{1}}$	$M_{2,\hat{2}}$	$\cdots$	$M_{2,\hat{C}}$	$M_{2,:}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C	$M_{C,\hat{1}}$	$M_{C,\hat{2}}$	$\cdots$	$M_{C,\hat{C}}$	$M_{C,:}$
Suma:	$M_{:,\hat{1}}$	$M_{:,\hat{2}}$	$\cdots$	$M_{:,\hat{C}}$	M

Normalización por filas: estimación empírica de $p(\hat{y} \mid y)$

Normalización por columnas: estimación empírica de $p(y \mid \hat{y})$

Normalización por filas y columnas: estimación empírica de $p(y, \hat{y})$

Análisis de una clase específica: se reduce a matriz binaria considerando el resto de clases como una única clase (negativa)

3 Problemas de regresión

Estados de la naturaleza y acciones: reales, $\mathcal{H}=\mathcal{A}=\mathcal{Y}=\mathbb{R}$

Pérdida L2 (ℓ_2, cuadrática o error cuadrático):

$$\ell_2(h,a)=(h-a)^2$$

Pérdida L1 (ℓ_1):

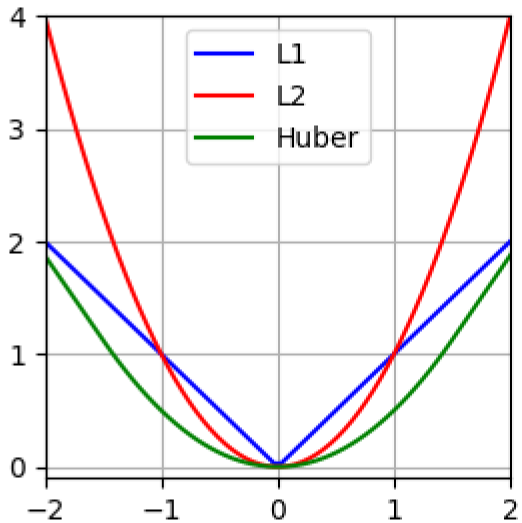
$$\ell_1(h,a)=|h-a|$$

Pérdida Huber: combina L1 y L2 con un parámetro $\delta\geq 0$

$$\ell_\delta(h,a)=\begin{cases}\frac{(h-a)^2}{2} & \text{si } |h-a|\leq\delta\\ \delta|h-a|-\frac{\delta^2}{2} & \text{si } |h-a|>\delta\end{cases}$$

Comparación gráfica: L2, L1 y Huber en función de la desviación de la verdad, $h - a$

In [1]:
```
import numpy as np; import matplotlib.pyplot as plt
e = np.linspace(-3.0, 3.0, 100) # e = h - a (error)
L1 = abs(e); L2 = np.square(e); delta = 1.5; i = abs(e) <= delta
Huber = (abs(e)<=delta) * 0.5*L2 + (abs(e)>delta) * delta*(L1-delta/2);
plt.figure(figsize=(3, 3)); plt.xlim((-2, 2)); plt.ylim((-0.1, 4)); plt.grid();
plt.plot(e, L1, 'b'); plt.plot(e, L2, 'r'); plt.plot(e, Huber, 'g'); plt.legend(['L1', 'L2', 'Huber']);
```



Observaciones que se derivan de la comparación gráfica:

- L1 penaliza **linealmente** las desviaciones de la verdad
- L2 penaliza **cuadráticamente** las desviaciones de la verdad, por lo que es **más sensible a outliers** que L1
- Huber representa un compromiso entre L1 y L2

Pérdida L2 esperada a posteriori:

$$R(a \mid \boldsymbol{x}) = \mathbb{E}[(h-a)^2 \mid \boldsymbol{x}] = \mathbb{E}[h^2 \mid \boldsymbol{x}] - 2a\mathbb{E}[h \mid \boldsymbol{x}] + a^2$$

Regresor de Bayes L2 o minimum mean squared error (MMSE): media a posteriori

$$\frac{\partial}{\partial a} R(a \mid \boldsymbol{x}) = -2\mathbb{E}[h \mid \boldsymbol{x}] + 2a = 0 \quad \Rightarrow \quad \pi(\boldsymbol{x}) = \mathbb{E}[h \mid \boldsymbol{x}] = \int h\, p(h \mid \boldsymbol{x})\, dh$$

Pérdida L1 esperada a posteriori:

$$R(a \mid \boldsymbol{x}) = \mathbb{E}[|h-a| \mid \boldsymbol{x}] = \int |h-a|\, p(h \mid \boldsymbol{x})\, dh = \int_{\infty}^{a} (a-h)\, p(h \mid \boldsymbol{x})\, dh + \int_{a}^{\infty} (h-a)\, p(h \mid \boldsymbol{x})\, dh$$

Regresor de Bayes L1: mediana a posteriori

$$a: \; P(h < a \mid \boldsymbol{x}) = P(h \geq a \mid \boldsymbol{x}) = 0.5$$

Pérdidas para $\mathbb{R}^D$: las pérdidas usuales para $\mathbb{R}$ pueden extenderse fácilmente a $\mathbb{R}^D$ y usarse para calcular los parámetros óptimos que debe devolver un estimador, la acción óptima que debe realizar un robot, etc.

4 Problemas de predicción probabilística

Estados de la naturaleza y acciones: distribuciones de probabilidad; $h = p(Y \mid \boldsymbol{x})$ y buscamos una $a = q(Y \mid \boldsymbol{x})$ que minimice $\mathbb{E}[\ell(p, q)]$ para un $\boldsymbol{x}$ dado

Función de pérdida: **divergencia de Kullback-Leibler (KL),** en función de la entropía de p, $\mathbb{H}(p)$, y la entropía cruzada entre p y q, $\mathbb{H}(p, q)$

$$\mathbb{KL}(p \| q) = -\mathbb{H}(p) + \mathbb{H}(p, q)$$

La minimización de KL equivale a minizar la entropía cruzada:

$$\begin{aligned} q^*(Y \mid \boldsymbol{x}) &= \operatorname*{argmin}_q \; \mathbb{KL}(p(Y \| \boldsymbol{x}), q(Y \mid \boldsymbol{x})) \\ &= \operatorname*{argmin}_q \; -\mathbb{H}(p) + \mathbb{H}(p(Y \mid \boldsymbol{x}), q(Y \mid \boldsymbol{x})) \\ &= \operatorname*{argmin}_q \; \mathbb{H}(p(Y \mid \boldsymbol{x}), q(Y \mid \boldsymbol{x})) \\ &= \operatorname*{argmin}_q \; -\sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}) \log q(y \mid \boldsymbol{x}) \end{aligned}$$

En clasificación equivale a usar la log-pérdida: si h es one-hot, $h = p(Y \mid \boldsymbol{x}) = \delta(Y = c)$

$$\mathbb{H}(\delta(Y = c), q) = -\sum_{y \in \mathcal{Y}} \delta(y = c) \log q(y \mid \boldsymbol{x}) = -\log q(c \mid \boldsymbol{x})$$

Regla de puntuación propia: pérdida $\ell(p, q)$ cuya minimización en q converge a p

Puntuación de Brier: regla propia menos sensible a eventos raros que la entropía cruzada

$$\ell(p, q) = \frac{1}{C} \sum_{c=1}^{C} \left(q(y = c \mid \boldsymbol{x}) - p(y = c \mid \boldsymbol{x})\right)^2$$