Loss Functions

Quadratic loss conditional expectation theorem

Prereq: ① $L(Y, f(x)) = (Y - f(x))^2$

② $E_{X,Y}[L(Y, f(x))] = E_{X,Y}(Y - f(x))^2$

Claim: f that minimizes ② is $f(x) = E_{Y|X}[Y \mid X]$

Proof:

① Assume that c is some constant $\in \mathbb{R}$. We will show that $c = \mu$ will minimize $E_Y[(Y-c)^2]$

$$\frac{\partial}{\partial c} E_Y[(Y-c)^2] = -E\left[2(Y-c)\right]$$

$$= -2E[Y] + 2c$$

Setting to 0:

$$-2E[Y] + 2c = 0$$

$$c = E[Y]$$

This is also a minimum b/c we differentiated a quadratic function that opens upwards.

② Now, assume $\hat{c}(x)$. From above argument, $\hat{c}(x) = E[Y \mid X = x]$ minimizes $E_{Y|X=x}[(Y - c(x))^2 \mid X = x]$.

③ By minimality, we can say that if $Z = z(x)$, $W = w(x)$, then $Z \geq W \Leftrightarrow \forall x,\ z(x) \geq w(x)$.

Then, from ②, we can say for any function $g(\cdot)$:

$$E_Y[(Y - g(x))^2 \mid X = x] \geq E_Y[(Y - \hat{c}(x))^2 \mid X = x]$$

$$\Rightarrow E_Y[(Y - g(x))^2] \geq E_Y[(Y - \hat{c}(x))^2]$$

④ Taking E_X on both sides & using law of iterated expectation:

$$E_X[E_Y[(Y - g(x))^2 \mid X = x]] \geq E_X[E_Y[(Y - \hat{c}(x))^2 \mid X = x]]$$

$$E_{X,Y}[(Y - g(x))^2] \geq E_{X,Y}[(Y - \hat{c}(x))^2]$$

⑤ This shows us that $\hat{c}(x) = E[Y \mid X]$ minimizes the risk function.

Absolute loss risk minimization

Prereq: ① $L(Y, f(X)) = |Y - f(X)|$

Claim: $f(\cdot)$ that minimizes $\mathbb{E}[L(Y, f(X))]$ is Median $(Y \mid X = x)$

Proof: ① Let $f_Y(y)$ be the density function of Y. Assume $f(x)$ is a constant c for now. Then:

$$\mathbb{E}[L(Y, f(c)] = \int_{-\infty}^{\infty} |y - c| f_Y(y)\, dy$$

$$= \underbrace{\int_{-\infty}^{c} (c-y) f_Y(y)\, dy}_{I_1} + \underbrace{\int_{c}^{\infty} (y-c) f_Y(y)\, dy}_{I_2}$$

Differentiating w/ respect to c:

$$\frac{d}{dc} \mathbb{E}[L(Y, f(c))] = \frac{d}{dc} I_1 + \frac{d}{dc} I_2$$

Applying Leibniz Rule which states:

$$\frac{d}{dc} \int_{a(c)}^{b(c)} f(y, c)\, dy = f(b(c), c) \cdot \frac{d b(c)}{dc} - f(a(c), c) \cdot \frac{d a(c)}{dc} + \int_{a(c)}^{b(c)} \frac{\partial f(y, c)}{\partial c}\, dy$$

$$\therefore \frac{d}{dc} I_1 = (c-c) f_Y(c) \cdot 1 - f(\infty, c) \cdot 0 + \int_{-\infty}^{c} f_Y(y)\, dy$$

$$= \int_{-\infty}^{c} f_Y(y)\, dy$$

$$\frac{d}{dc} I_2 = (\infty - c) f_Y(\infty) \cdot 0 - (c-c) f_Y(c) \cdot 1 - \int_{c}^{\infty} f_Y(y)\, dy$$

$$= -\int_{c}^{\infty} f_Y(y)\, dy$$

Thus:

$$\frac{d}{dc} \mathbb{E}[L(Y, f(c))] = \frac{d}{dc} I_1 + \frac{d}{dc} I_2$$

$$= \int_{-\infty}^{c} f_Y(y) - \int_{c}^{\infty} f_Y(y)\, dy$$

$$= F_Y(c) - (1 - F_Y(c))$$

$$= -1 + 2 F_Y(c)$$

Setting to 0:

$$F_Y(c) = 1/2 \Rightarrow \text{Median}$$

Thus, $f(\cdot)$ that minimizes this expectation is Median (Y).

② If c now becomes $\hat{c}(x)$, we have:

$$\mathbb{E}[|Y-\hat{c}(x)| \mid X=x]$$

We know that $c(x) = \text{Median}(Y \mid X=x)$ still holds since $\hat{c}(x)$ is treated like a constant.

③ From minimality, we know that for any function $g(\cdot)$,

$$\mathbb{E}[|Y-g(X)| \mid X=x] \geq \mathbb{E}[|Y-\hat{c}(x)| \mid X=x]$$

So:

$$\mathbb{E}[|Y-g(X)| \mid X] \geq \mathbb{E}[|Y-\hat{c}(X)| \mid X]$$

④ Taking iterated expectations:

$$\mathbb{E}_X[\mathbb{E}[|Y-g(X)| \mid X]] \geq \mathbb{E}_X[\mathbb{E}[|Y-\hat{c}(X)| \mid X]]$$

$$\mathbb{E}_{X,Y}[|Y-g(X)|] \geq \mathbb{E}_{X,Y}[|Y-\hat{c}(X)|]$$

⑤ This shows that $\hat{c}(X) = \text{Median}(Y \mid X)$ is the minimizer for Abs. Loss.

REGRESSION REVIEW

Proof of $\hat{\beta}$ estimation

Claim: If $Y = X\beta + \varepsilon$, $\hat{\beta}$ that best mimizes square loss is $\hat{\beta} = (X'X)^{-1}X'Y$

① Model specification:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim MVN(0, \sigma^2 I).$$

② Squared error loss function:

$$(Y-X\beta)'(Y-X\beta)$$

③ Differentiate:

$$\frac{\partial}{\partial\beta}[(Y-X\beta)'(Y-X\beta)] = \frac{\partial}{\partial\beta}[(Y'-\beta'X')(Y-X\beta)]$$

$$= \frac{\partial}{\partial\beta}[Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta]$$

$$= -(Y'X)' - (X'Y) + (X'X + (X'X)')\beta$$

$$= -2X'Y + 2X'X\beta$$

④ Set to 0

$$-2X'Y + 2X'X\hat{\beta} = 0$$

$$X'X\hat{\beta} = X'Y$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Idempotency of hat matrix

Claim: $H = X(X'X)^{-1}X'$. $H = H^2$

Proof:

$$H^2 = X(X'X)^{-1}X' \cdot X(X'X)^{-1}X'$$

$$= X\underbrace{\cancel{(X'X)^{-1}X'X}}_{I} \cdot (X'X)^{-1}X'$$

$$= X(X'X)^{-1}X'$$

$$= H$$

Orthogonality of $\hat{r}$ & $\hat{Y}$

Claim: $\hat{r} \cdot \hat{Y} = 0$

Proof: $\hat{r} = Y - \hat{Y} = (I - H)Y$

$\hat{Y} = HY$

$$\therefore \hat{r} \cdot \hat{Y} = \hat{r}'\hat{Y}$$

$$= (Y - HY)'(HY)$$

$$= Y'HY - Y'H'HY$$

By symmetry of H

$$= Y'HY - H \cdot Y'HHY$$

By idempotency of H

$$= Y'HY - Y'HY$$

$$= 0$$

PREDICTION W/ LINEAR REGRESSION

Prediction interval theorem

Claim: Let (y_0, x_0) ∉ training set. Let us define a point forecast of y_0 at $X = x_0$ as $\hat{\mu} = x_0' \hat{\beta}$.

Then, $100(1-\alpha)\%$ prediction interval is:

$$\hat{\mu} \pm c_\alpha \hat{\sigma} \sqrt{1 + x_0'(X'X)^{-1} x_0'}$$

Proof: We know that:

$$Y_0 - x_0' \hat{\beta} \sim N(0, \sigma^2(1 + x_0'(X'X)^{-1} x_0)$$

Then:

$$T := \frac{Y_0 - x_0' \hat{\beta}}{\hat{\sigma} \sqrt{1 + x_0'(X'X)^{-1} x_0'}} \sim t_{n-p-1}$$

For a confidence interval:

$$P(-c_\alpha \leq T \leq c_\alpha) = 1 - \frac{\alpha}{2}$$

$$\therefore \hat{\mu} \pm c_\alpha \hat{\sigma} \sqrt{1 + x_0'(X'X)^{-1} x_0'} \quad \text{is } 100(1-\alpha)\% \text{ prediction interval}$$

BIAS-VARIANCE TRADEOFF

Decomposition

Claim: We have a training set T & an estimate of β, $\hat{\beta} = \hat{\beta}(T)$.
We calculate the forecasting error for new x_0, $\hat{y}_0 = x_0' \hat{\beta}$.
Then:

$$MSE(x_0) = \text{Bias}^2(y_0) + \text{Var}_T(y_0)$$

Proof:

$$MSE(x_0) := E_T[(y_0 - \hat{y}_0)^2]$$

$$= E_T[(y_0 - E_T(\hat{y}_0) + E_T(\hat{y}_0) - \hat{y}_0)^2]$$

$$= \underbrace{(y_0 - E_T(\hat{y}_0))^2}_{\text{Bias}^2} + \underbrace{E_T((\hat{y}_0 - E_T(\hat{y}_0))^2)}_{\text{Variance}}$$

Prediction error decomposition

Claim: $Y = f(X) + \epsilon$, $Var(\epsilon) = \sigma_\epsilon^2$. For some input $X_0 = x_0$, output is $\hat{f}(x_0)$

Then: $Err(x_0) = \sigma_\varepsilon^2 + Bias^2 + Variance$

Proof:

$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 \mid X_0 = x_0]$$

$$= E[Y^2 + \hat{f}(x_0)^2 - 2Y\hat{f}(x_0) \mid X_0 = x_0]$$

$$= E[Y^2 \mid X_0 = x_0] + E[\hat{f}(x_0)^2 \mid X_0 = x_0] - 2E[Y\hat{f}(x_0) \mid X_0 = x_0]$$

$$= Var(Y \mid X_0 = x_0) + E[Y \mid X_0 = x_0]^2 + \| - 2E[(f(x_0) + \underbrace{\varepsilon)\hat{f}(x_0)}_{E[\varepsilon \ldots] = 0} \mid X_0 = x_0]$$

constant

$$= \sigma_\varepsilon^2 + E[f(x_0) \mid X_0 = x_0]^2 + \| - 2E[f(x_0)\hat{f}(x_0) \mid X_0 = x_0]$$

$$= \sigma_\varepsilon^2 + f(x_0)^2 + E[\hat{f}(x_0)^2 \mid X_0 = x_0] - 2E[f(x_0)\hat{f}(x_0) \mid X_0 = x_0]$$

$$= \sigma_\varepsilon^2 + E[(f(x_0) - \hat{f}(x_0))^2 \mid X_0 = x_0] \quad \pm E[\hat{f}(x_0)]$$

$$= \sigma_\varepsilon^2 + E[(f(x_0) - E[\hat{f}(x_0)])^2 \mid X_0 = x_0] + E[(E(\hat{f}(x_0)) - \hat{f}(x_0))^2 \mid X_0 = x_0]$$

$$= \sigma_\varepsilon^2 + Bias^2 + Variance$$

REGULARIZATION METHODS

Ridge regression closed form derivation

① Minimization:

$$\underset{\beta_0, \vec{\beta}}{\operatorname{argmin}}\ RSS(\lambda) = \underset{\beta_0, \vec{\beta}}{\operatorname{argmin}}\ (y - \beta_0\vec{1} - X\beta)'(y - \beta_0\vec{1} - X\beta) + \lambda\beta'\beta$$

② Derivative.

$$RSS(\lambda) = (y' - \vec{1}'\beta_0' - \beta'X')(y - \beta_0 1 - X\beta) + \lambda\beta'\beta$$

$$= y'y - y'\beta_0\vec{1} - y'X\beta - \vec{1}'\beta_0'y - \vec{1}'\beta_0'\beta_0 1 - \vec{1}'\beta_0'X\beta - \beta'X'y$$
$$+ \beta'X'\beta_0\vec{1} + \beta'X'X\beta + \lambda\beta'\beta$$

$$\frac{\partial}{\partial\beta} RSS(\lambda) = -(Y'X)' - (\vec{1}'\beta_0'X)' - (X'Y) + (X'\beta_0\vec{1}) + (X'X + X'X)\beta + 2\lambda\beta$$

$$= -2X'Y + 2X'X\beta + 2\lambda\beta$$

(3) Setting to 0:

$$-2X'Y + 2X'X\hat{\beta} + 2\lambda\hat{\beta} = 0$$
$$-X'Y + X'X\hat{\beta} + \lambda\hat{\beta} = 0$$
$$(X'X + \lambda I)\hat{\beta} = X'Y$$
$$\hat{\beta} = (X'X + \lambda I)^{-1}X'Y$$

STATIONARY PROCESSES

Strictly stationary & weakly stationary relationship

Claim: Strictly stationary process $\Rightarrow$ weakly stationary process

Proof: Assume that $\{X_t\}$ is a strictly stationary process. We will show that it is weakly stationary.

(1) $E[X_t^2] < \infty$ and indep. of t by defn.

(2) $E[X_t]$: Since $F_{X_1}(x) = \dots = F_{X_n}(x)$ $\forall x$, the mean function $\mu_X(t)$ must be constant & indep. of t.

(3) $\gamma(h)$:

$$\gamma(h) = Cov(X_t, X_{t+h})$$
$$= E[X_t X_{t+h}] - E[X_t]E[X_{t+h}]$$
$$= E[X_1 X_{h+1}] - E[X_1]E[X_{h+1}]$$

Since $\{X_t\}$ is strictly stationary, we know that $F_{X_1, X_2}(x_1, x_2) = F_{X_{1+h}, X_{2+h}}(x_1, x_2)$ $\forall k, l$. Thus:

$$E[X_t X_{t+h}] = E[X_1 X_{h+1}]$$

We also know that all univariate distributions will be the same as $\forall h$, $F_{X_1}(x_1) = F_{X_{1+h}}(x_1)$. Thus:

$$E[X_t] = E[X_1]$$
$$E[X_{t+h}] = E[X_{1+h}]$$

Since all components of $Cov(X_t, X_{t+h})$ is independent of t, the covariance will also be indep. of t.

$\therefore$ (1), (2), (3) $\Rightarrow$ weakly stationary

SMOOTHING METHODS

Exponential Smoothing Closed Form

Claim: The closed form solution of an exponentially smoothed average is

$$\hat{m}_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{t-1} X_1 \qquad (t \geq 2)$$

Proof: The recursive defn. is:

$$\hat{m}_t = \begin{cases} \alpha X_t + (1-\alpha)\hat{m}_{t-1} & , t > 1 \\ X_1 & , t = 1 \end{cases}$$

We will use induction to prove that the closed form (hereby called P_t) is equivalent to the recursive form.

For $t = 2$:

$$P_2 = \sum_{j=0}^{2-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{2-1} X_1$$

$$= \alpha X_2 + (1-\alpha) X_1$$

This is equivalent to the recursive version

Assume P_t is equivalent to $\hat{m}_t$ $\forall\ 2 \leq t \leq k$.

Consider P_{t+1}:

$$P_{t+1} = \sum_{j=0}^{t+1-2} \alpha(1-\alpha)^j X_{t+1-j} + (1-\alpha)^{t+1-1} X_1$$

$$= \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t+1-j} + (1-\alpha)^t X_1$$

$$=$$

STATIONARY LINEAR PROCESSES

Stationary MA(q) process

Claim: MA(q) process is stationary

Proof:

Let X_t be an MA(q) process w/ the following process

$$X_t = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q} \qquad \text{where } Z_i \sim WN(0, \sigma^2)$$

① Expectation:

$$E[X_t] = E[Z_t] + \theta_1 E[Z_{t-1}] + \dots + \theta_q E[Z_{t-q}]$$
$$= 0 \Rightarrow \text{indep. of } t$$

② Variance:

$$Var[X_t] = \overset{Var[Z_t]}{\wedge} \theta_1^2 Var[Z_{t-1}] + \dots + \theta_q^2 Var[Z_{t-q}] + \underbrace{\theta_1 \theta_2 Cov(Z_t, Z_{t-1}) \dots}_{0 \text{ b/c WN}}$$
$$= \theta_1^2 \sigma^2 + \dots + \theta_q^2 \sigma^2 + \sigma^2$$
$$= \sigma^2 (1 + \theta_1^2 + \dots + \theta_q^2)$$

Variance is finite & indep. of t.

③ ACVF

$$\gamma(h) = Cov(X_t, X_{t+h}) \qquad \theta_2 Z_{t-2+h}$$
$$= Cov(Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q},\ Z_{t+h} + \theta_1 Z_{t-1+h} + \dots + \theta_q Z_{t-q+h})$$

If $h = 1$:

$$\gamma(h) = \theta_0 \theta_1 \sigma^2 + \theta_1 \theta_2 \sigma^2 + \dots + \theta_{q-1} \theta_q \sigma^2$$

If $0 < h < q$

$$\gamma(h) = \theta_0 \theta_h \sigma^2 + \dots + \theta_{q-h} \theta_q \sigma^2$$

If $h = q$:

$$\gamma(h) = \theta_0 \theta_q \sigma^2$$

$h > q$:

$$\gamma(h) = 0$$

Since $\gamma(h)$ is not dependent on t, ①, ②, ③ are satisfied $\Rightarrow$ MA(q) is stationary.

Best linear predictor

Claim: $P_n X_{n+h} = a_0 + a_1 X_n + \dots + a_n X_1$ is the best linear predictor in terms of MSE if:

a) $a_0 = \mu(1 - \sum_{i=1}^{n} a_i)$

b) $a_1, \dots, a_n$ satisfy $\Gamma_n \vec{a}_n = \gamma_n(h)$

Proof: ① Minimization

$$S = \underset{a_0, \dots, a_n}{\operatorname{argmin}} \quad \mathbb{E}\left[(X_{n+h} - P_n X_{n+h})^2\right]$$

② $\dfrac{\partial S}{\partial a_j}$

Eq 1: $$\frac{\partial S}{\partial a_0} = \frac{\partial}{\partial a_0} \mathbb{E}\left[(X_{n+h} - a_0 - a_1 x_n \dots - a_n x_1)^2\right] = 0$$

$$= -2\,\mathbb{E}\left[X_{n+h} - a_0 - a_1 x_1 \dots - a_n x_1\right] = 0$$

$$\mu - a_0 - a_1\mu \dots - a_n\mu = 0$$

$$a_0 = \mu + a_1\mu + \dots + a_n\mu$$

$$= \mu\left(1 + \sum_{i=1}^{n} a_i\right)$$

Eq 2: For $j = 1, 2, \dots, n$

$$\frac{\partial S}{\partial a_j} = \frac{\partial}{\partial a_j} \mathbb{E}\left[(X_{n+h} - a_0 - a_1 X_n - \dots a_n X_1)^2\right] = 0$$

$$\Rightarrow \quad -2\,\mathbb{E}\left[X_{n+1-j}\left(X_{n+h} - a_0 - a_1 X_n - \dots - a_n X_1\right)\right] = 0$$

$$\Rightarrow \quad \mathbb{E}[X_{n+1-j} X_{n+h}] - a_0 \cancel{\mathbb{E}[X_{n+1-j}]}^{\mu} - a_1 \mathbb{E}[X_n X_{n+1-j}] \dots = 0$$

$$\Rightarrow \quad \mathbb{E}[X_{n+1-j} X_{n+h}] - \mu^2\left(1 - \sum_{i=1}^{n} a_i\right) - \sum_{i=1}^{n} a_i\, \mathbb{E}[X_{n+1-j} X_{n+1-i}] = 0$$

$$\Rightarrow \left(\mathbb{E}[X_{n+1-j} X_{n+h}] - \mu^2\right) + \mu^2 \sum_{i=1}^{n} a_i - \quad \text{''} \quad = 0$$

$$\Rightarrow \quad \gamma(h+j-1) - \sum_{i=1}^{n} a_i\left(\mathbb{E}[X_{n+1-j} X_{n+1-i}] - \mu^2\right) = 0$$

$$\Rightarrow \quad \gamma(h+j-1) - \sum_{i=1}^{n} a_i\, \gamma(i-j) = 0$$

$\therefore$ For diff values of j

$$\gamma_n(h) = \Gamma_n \vec{a_n}$$

Best linear predictor - MSE

Claim: $\mathbb{E}\left[(X_{n+h} - P_n X_{n+h})^2\right] = \gamma(0) - a_n' \gamma_n(h)$

Proof:

$$\mathbb{E}\left[(X_{n+h} - P_n X_{n+h})^2\right] = \mathbb{E}\left[\left((X_{n+h} - \mu) - a'(\vec{X} - \mu)\right)^2\right]$$

$$= \mathbb{E}\left[(X_{n+h}-\mu)^2 - 2(X_{n+h}-\mu)\, a'(\vec{X}-\mu) + (a'(\vec{X}-\mu))^2 \right]$$

?? I don't know how to prove this.

Box - Jenkins Models

ACVF of ARMA(p, q)

Claim: ACVF of stationary & causal ARMA(p, q) process is

$$\gamma(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}$$

Proof: If X_t is an ARMA(p, q) process, then it can be written as:

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

We can now evaluate:

$$\gamma(h) = Cov(X_t, X_{t+h})$$

$$= \mathbb{E}[X_t X_{t+h}] - \cancel{E[X_t]}^{\to 0} E[X_{t+h}]$$

$$= \mathbb{E}\left[\sum_{j=0}^{\infty} \psi_j Z_{t-j} \sum_{i=0}^{\infty} \psi_i Z_{t+h-i} \right]$$

Reindex

$$= \sum_{j=0}^{\infty} \psi_j \psi_{j+h}\, E[Z_j^2]$$

$$= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}$$