# Quiz 2

# Gaussian Discriminant Analysis

Austin Welch

- Generative algorithm
- models $p(x|y)$ & $p(y)$ (likelihood, prior)
- can get $p(y|x)$ from Bayes' Rule (Posterior)
- assumes $p(x|y)$ is Gaussian

$$h_{MAP}(x) = \underset{y=1,\dots,m}{\arg\max}\ \underbrace{P(y|x,\theta)}_{\text{posterior}} = h_{GDA}(x)$$

$$= \underset{y}{\arg\min}\ [-\ln \underbrace{P(x|y,\theta)}_{\text{likelihood}} - \ln \underbrace{P(y)}_{\text{prior}}]$$

remember $\mu_y$ is a vector!

$$= \underset{y}{\arg\min}\left[\underbrace{\frac{1}{2}(x-\mu_y)^T\Sigma_y^{-1}(x-\mu_y)}_{\text{class-dependent quadratic in } x} + \underbrace{\frac{1}{2}\ln\det(\Sigma_y) - \ln P(y)}_{\text{class-dependent scalar offset}}\right]$$

multivariate Gaussian

RV $z \sim \mathcal{N}(\vec{\mu}, \Sigma)$

$$P(z) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\right\}$$

($|\Sigma|^{1/2} \leftarrow \det(\Sigma)$)

$$\Sigma = E[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T]$$

eg. for $d=2$, if $\Sigma$ = diagonal, variables are uncorrelated

ex. $\Sigma = \begin{bmatrix}1&0\\0&1\end{bmatrix}$

Diagonals shrink/stretch dimensions (shaper/flatter peak)

if $\Sigma$ not diagonal, variables become correlated which has effect of flattening out the Gaussian in $x=y$ or $x=-y$ dimension

ex. $\Sigma = \begin{bmatrix}1&.8\\.8&1\end{bmatrix}$

($\Sigma$ is symmetric and positive-semi-definite)

ex. $\Sigma = \begin{bmatrix}1&-.8\\-.8&1\end{bmatrix}$

Say $P(y) = \phi^y(1-\phi)^{1-y}$ (Bernoulli)

then, for $d=2$,

$$P(x|y=0)\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)$$

$$P(x|y=1) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

And,

$$\ell(\phi,\mu_0,\mu_1,\Sigma) = \log\prod_{i=1}^{m}P(x^{(i)},y^{(i)}) \quad \text{(Joint likelihood)}$$

$$= \log\prod_{i=1}^{m}P(x^{(i)},y^{(i)})\,P(y^{(i)})$$

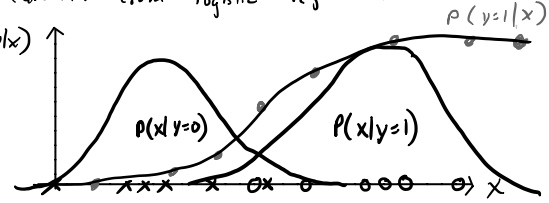
**QDA in words**: map $x$ to the class whose mean vector $\mu_y$ is closest to it. Here, closeness is measured by a suitably weighted and offset-adjusted distance between $x$ & $\mu_y$.

Linear boundary if $\Sigma_y$'s of adjacent classes are equal. (otherwise parabolic)

Note: $\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$ is the Mahalanobis distance ($\sigma$-D std's from mean)

## Connection to logistic regression

When you assume $p(x|y)$ is Gaussian in GDA, if you plot the probabilities of $p(y|x)$, you end up with the sigmoid function from logistic regression.




But, GDA will choose different position & steepness for the sigmoid than logistic regression.

$x|y \sim$ Gaussian (stronger assumption) $\Rightarrow$ logistic posterior for $p(y=1|x)$ $\not\Leftarrow$

**LDA**: $\Sigma_y = \Sigma$ for all $y$

Then quadratic rule becomes linear rule:

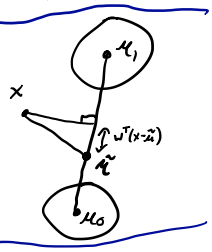
$$h_{LDA}(x) = \underset{y}{\arg\max}\left[\underbrace{\beta_y^T x}_{\text{projection}} + \underbrace{\gamma_y}_{\text{offset}}\right]$$

**2-class LDA**

$h_{LDA}(x) = 1(g(x) > \tau)$

$g(x) = w^T x$

$\tau = w^T\tilde{\mu}$




$b^T x$: inner product of 2 vectors is proportional to the correlation between their components

**LDA in Words**: Map $x$ to the class whose mean vector $\mu_y$ is maximally correlated with it. Here, the correlation is measured by a suitably weighted and offset-adjusted correlation between the components of $\mu_y$ and $x$.

Posterior PMF:

$$p_{LDA}(y|x,\theta) = \frac{e^{\beta_y^T x+\gamma_y}}{\sum_{j=1}^{m}e^{\beta_j^T x+\gamma_j}} = \text{softmax}(y|\eta)$$

where $\eta = (\beta_1^T x+\gamma_1,\dots,\beta_m^T x+\gamma_m) = (\eta_1,\dots,\eta_m)^T$

**Gen. Alg.** ⊕/⊖

⊕ requires less data (because of stronger assumption)

⊖ less robust (eg fam. $p(x|y) \Rightarrow$ logistic $p(y|x)$)

**Unbiased Estimates**

QDA: $\hat{\Sigma}_y \to \frac{n_y}{n_y-1}\hat{\Sigma}_y$

LDA: $\hat{\Sigma} \to \frac{n}{n-m}\hat{\Sigma}$

**Empirical Parameter Estimates**

$$n_y = \sum_{j=1}^{n}1(y_j=y)$$

$$\hat{P}(y) = n_y/n, \quad \hat{\mu}_y = \frac{\sum_{j\in\{1,\dots,n\}:y_j=y}x_j}{n_y}$$

QDA $\hat{\Sigma}_y = \frac{1}{n_y}\sum_{j\in\{1,\dots,n\}:y_j=y}(x_j-\hat{\mu}_y)(x_j-\hat{\mu}_y)^T$

LDA $\hat{\Sigma} = \sum_{y=1}^{m}\hat{P}(y)\hat{\Sigma}_y$ SAME AS ML ESTIMATES!

**RDA key idea**

A: non-invertible Cov mat (positive semi-definite)

B: symmetric, Positive-definite (so B is invertible)

Then for $C = A+B$, C is symmetric Pos. def. and therefore, invertible.

Special case with $\lambda=1$: $\hat{\Sigma}_{reg} = \text{diag}(\hat{\Sigma}_{ML})$

Equivalent to assuming that all $d$ components are independent. $\Rightarrow$ "Gaussian Naive Bayes"

Need for Regularization:

Overfitting: Complexity of model is greater than the amount of data, & may can still be ill-conditioned when $d < n$

Reg strategy #1: $\hat{\Sigma}_{reg} = \lambda I_d + (1-\lambda)(\hat{\Sigma}_{ML})$

Reg strat. #2: $\hat{\Sigma}_{reg} = \lambda\,\text{diag}(\hat{\Sigma}_{ML}) + (1-\lambda)(\hat{\Sigma}_{ML})$

$\theta$: parameters
$x$: evidence
$p(x)$: evidence
$p(\theta)$: priors
$p(x|\theta)$: likelihood
$p(\theta|x)$: posterior

**Generative**
$p(x, \theta) = p(x|\theta)\,p(\theta)$

**discriminative**
$p(\theta|x)$ directly

Don't know $p(x|\theta)$, $p(\theta)$

Posterior $\propto$ likelihood $\cdot$ prior (Proportional, NOT equal)
$p(\theta|x) \propto p(x|\theta)\,p(\theta)$

---

If $X = Y + Z$ and $Y, Z$ are independent

Gaussians: conditional mean
$Y|X=x \sim N(E[Y|X=x], \mathrm{Cov}(Y, X=x))$
$E[Y|X=x] = E[Y] + \mathrm{Cov}(Y,X)\,\mathrm{Cov}(X)^{-1}(x - E[X])$

Cond. cov $\mathrm{Cov}(Y|X=x) = \mathrm{Cov}(Y) - \mathrm{Cov}(Y,X)\,\mathrm{Cov}(X)^{-1}\,\mathrm{Cov}(X,Y)$

in this problem, then $\mathrm{Cov}(Y, Y+Z) = \mathrm{Cov}(Y,Y) + \cancel{\mathrm{Cov}(Y,Z)}$

and likelihood $P(x|y) = N(y, \sigma_2^2)$

---

Risk of decision rule is exp. val of loss

$R(h) = E[\ell(X,Y,h)]$ Risk is expected loss

$h_{emp}(h) = \frac{1}{n}\sum_{j=1}^{n} \ell(x_j, y_j, h) \xrightarrow{n\to\infty} E[\ell(X,Y,h)]$

$= R(h)$ (needs $p(x,y)$)

$E[\ell(h(x), y)]$

$= \int \ell(h(x), y)\, p(x,y)\, dx\, dy$

for ex. $\ell = (y - h(x))^2$

---

$R_{Bayes}$: min possible risk over all decision rules when $p(x,y)$ is known

$h_{Bayes}$: Best decision rule when $p(x,y)$ is known.

$R(h_{Bayes}) = R_{Bayes}$ (Bayes decision rule attains Bayes Risk)

Best in-family decision rule minimizes the in-family posterior risk.

$h_H(x) = \underset{h\in H}{\mathrm{argmin}}\, E[\ell(X,Y,h) \mid X=x]$ Posterior risk given $X$

$h_{Bayes}(x) = \underset{y}{\mathrm{argmin}}\, E[\ell(Y, \hat{y}) \mid X=x]$

---

$z = \frac{x-\mu_x}{\sigma_x}$ $\quad Q(z)$

Gaussian
$X \sim N(\mu, \sigma^2)$
$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

---

independence
$P(X \cap Y) = P(X)P(Y)$
$P(X|Y) = P(X)$
cond. ind.
$f_{X2|Y}(x_2|y) = f_{X|Y}(x|y)\, f_{2|Y}(z|y)$
indep. $\Rightarrow$ uncorrelated

---

$\mathrm{cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$
$= E[XY] - E[X]E[Y]$

Cov 0 $\to$ uncorrelated

$\mathrm{Var}(X) = E[X^2] - E[X]^2$

$\mathrm{Cov}(X,X) = \mathrm{var}(X)$

$\mathrm{Cov}(Y, Y+Z) = E[Y(Y+Z)] - E[Y]E[Y+Z]$
$= E[Y^2] + E[YZ] - E[Y](E[Y] + E[Z])$
$= \underbrace{E[Y^2] - E[Y]^2}_{\mathrm{Cov}(Y,Y)} + \underbrace{E[YZ] - E[Y]E[Z]}_{\mathrm{Cov}(Y,Z)}$

conditional expectation
$E[X|Y=y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\, dx$

---

If $X$ & $Y$ are Gaussians:
$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right)$
known that: $X|Y \sim N\left(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(Y-\mu_y),\ \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)$
So, conditional cov. mat. $\mathrm{Var}(X|Y) = \Sigma_x - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$ is constant and doesn't depend on $Y$. Therefore, $E[\mathrm{var}(X|Y)] = \mathrm{var}(X|Y)$

---

## Frequentist

$\hat{\theta}_{ML} = \underset{\theta}{\mathrm{argmax}}\, p(D|\theta)$
$= \underset{\theta}{\mathrm{argmax}} \left[\prod_{j=1}^{n} P(x_j, y_j|\theta)\right]$
$= \underset{\theta}{\mathrm{argmin}} \left[\frac{1}{n}\sum_{j=1}^{n} -\ln P(x_j, y_j|\theta)\right]$

## Bayesian

$\hat{\theta}_{MAP} = \underset{\theta}{\mathrm{argmax}}\, P(\theta|D)$
$= \underset{\theta}{\mathrm{argmax}} \left[P(D|\theta)\pi(\theta)\right]$
$= \underset{\theta}{\mathrm{argmin}} \left[\frac{1}{n}\sum_{j=1}^{n} -\ln P(x_j, y_j|\theta) - \frac{1}{n}\ln \pi(\theta)\right]$

---

**MAP** $h_{MAP}(x) = \text{Post. mode} = \underset{y}{\mathrm{argmax}}\, P(y|x)$
$= \underset{y}{\mathrm{argmax}}\, p(x,y) = \underset{y}{\mathrm{argmax}}\, p(x|y)\,p(y)$
$= \underset{y}{\mathrm{argmin}} -\ln p(x|y) - \ln p(y)$

**ML** $h_{ML}(x) = \underset{y}{\mathrm{argmax}}\, p(x|y)$, If $Y \sim \mathrm{Uni}(Y)$, $h_{MAP}(x) = h_{ML}(x)$

---

**MMSE** $\ell(x,y,h) = \|y - h(x)\|^2$
$h_{Bayes}(x) = h_{MMSE}(x) =$ posterior mean $= E[Y|X=x]$
Bayes risk for MMSE is $E[\|Y - h(x)\|^2 \mid X=x]$ (expected val of loss fn)

**MMAE** $\ell(x,y,h) = |y - h(x)|$
$h_{Bayes} = h_{MMAE} =$ Posterior median $= \int_{-\infty}^{h_{MMAE}(x)} p(y|x)\, dy$
$R_{Bayes} = E[|Y - \hat{y}| \mid X=x]$

More computationally

---

**Structural risk minimization**

$h^{sam}(d, \lambda) = \underset{h\in H}{\mathrm{argmin}} \left[\underbrace{\frac{1}{n}\sum_{j=1}^{n} \ell(x_j, y_j, h)}_{\text{Empirical risk (data term)}} + \frac{\lambda}{n}\underbrace{\text{complexity}(h)}_{\text{structure penalty (prior beliefs)}}\right]$

$d$: tuning, complexity controlling parameters; $\lambda$

---

## Regression

- Supervised learning
- label = quantity
- $Y \in \mathcal{Y} = \mathbb{R}$
- $\ell(x,y,h) := (y - h(x))^2$
- Risk $= E[(y - h(x))^2] = MSE$

$h_{Bayes}(x) = h_{MMSE}(x) = E[Y|X=x]$

### Ordinary Least Squares Regression

aka. linear least squares Reg.

$x \in \mathcal{X} = \mathbb{R}^d$ $\quad h(x) = w^T x + b$
$\theta = \begin{pmatrix} w \\ b \end{pmatrix} \in \mathbb{R}^{d+1}$

$\hat{\theta}_{OLS} = (\hat{w}_{OLS}, \hat{b}_{OLS})^T = \underset{(w,b)}{\mathrm{argmin}} \sum_{j=1}^{n} (y_j - w^T x_j - b)^2$

**Solution**
$\hat{w}_{OLS} = (\hat{\Sigma}_x)^{-1}\hat{\Sigma}_{xy}$
$\hat{b}_{OLS} = \hat{\mu}_y - \underbrace{\hat{\Sigma}_{yx}(\hat{\Sigma}_x)^{-1}}_{w_{OLS}^T}\hat{\mu}_x$

---

**Equivalent Expressions**

$\hat{\mu}_x = \frac{1}{n}[x_1 \cdots x_n]\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n}\mathcal{X} 1_n$

$\hat{\mu}_y = \frac{1}{n}\mathcal{Y} 1_n$

$\hat{\Sigma}_x = \frac{1}{n}\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T$, $\tilde{\mathcal{X}} = \cdots = \mathcal{X} B$ ($d \times n$) ($n \times n$)

$\hat{\Sigma}_{xy} = \frac{1}{n}\tilde{\mathcal{X}}\tilde{\mathcal{Y}}^T$

$\tilde{\mathcal{Y}} = [y_1 - \hat{\mu}_y, \ldots, y_n - \hat{\mu}_y] = \mathcal{Y} B$

So, $\hat{w}_{OLS} = (\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T)^{-1}\tilde{\mathcal{X}}\tilde{\mathcal{Y}}^T$

$\hat{b}_{OLS} = \frac{1}{n}\mathcal{Y} 1_n - (\hat{w}_{OLS})^T \frac{1}{n}\mathcal{X} 1_n$

---

**Solution:**

1) for any $w$, best $b$ choice is the one which minimizes the first (likelihood) term which is the same as in OLS
$\Rightarrow$ $b_{best}(w) = \hat{\mu}_y - w^T\hat{\mu}_x$ the cost function (in terms of only $w$) becomes
$\sum_{j=1}^{n} (\tilde{y}_j - w^T\tilde{x}_j)^2 + \lambda\|w\|_2^2$

1) for any $w$, best $b$ is given by $(\hat{\mu}_y - w^T\hat{\mu}_x)$ ←

2) $\hat{w}_{ridge} = (\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T + \lambda I_x)^{-1}\tilde{\mathcal{X}}\tilde{\mathcal{Y}}^T$

---

$\hat{\theta}_{OLS} = \begin{pmatrix} \hat{w}_{OLS} \\ \hat{b}_{OLS} \end{pmatrix} = \underbrace{(\mathcal{X}_{ext}\mathcal{X}_{ext}^T)^{-1}}_{(d+1)\times n}\underbrace{\mathcal{X}_{ext}}_{(d+1)\times n}\underbrace{\mathcal{Y}^T}_{n\times 1}$

ex. $d \times n = 1 \times 50$ (HW)

$x_j \to \begin{pmatrix} x_j \\ 1 \end{pmatrix}$ $\quad (d+1)\times 1$

then $\mathcal{X} \to \mathcal{X}_{ext} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & & 1 \end{bmatrix}$

$\mathcal{Y} = (y_1 \cdots y_n)$

$\mathcal{X} = (x_1 \cdots x_n)$

---

Whenever you see an optimization problem which looks like:
$\underset{w}{\mathrm{argmin}} \sum_{j=1}^{n} (\tilde{y}_j - w^T\tilde{x}_j)^2$

$\hat{w}_{opt} = (\tilde{\mathcal{X}}\tilde{\mathcal{X}}^T)^{-1}\tilde{\mathcal{X}}\tilde{\mathcal{Y}}^T$

where
$\tilde{\mathcal{X}} = (\tilde{x}_1 \cdots \tilde{x}_n)$
$\tilde{\mathcal{Y}} = (\tilde{y}_1 \cdots \tilde{y}_n)$

---

**Ridge Regression** ($l_2$-regularized)

$(\hat{w}_{ridge}, \hat{b}_{ridge}) := \underset{(w,b)}{\mathrm{argmin}} \left[\sum_{j=1}^{n} (y_j - w^T x_j - b)^T + \lambda\|w\|^2\right]$

N.L.L. $\quad$ $-\ln \pi(w)$ Prior

$y_j = w^T x_j + b + \text{noise}$ $N(0,1)$ $\quad$ $w \sim N(0, \frac{1}{2\lambda})$

This is equivalent to Bayesian Discriminative Learning with
$p(y|w,b,x) = N(w^T x + b, \frac{1}{2})$
$\pi(w) \sim N(0, \frac{1}{2\lambda})$

---

**Regularization:**
Penalize complexity
- solve ill-posed probs.
- reduce overfitting

# Naive Bayes

- Generative algorithm
- motivating example: spam classifier
- $d$ features in each feature vector
- NB assumption: all features are conditionally independent given the class label.

$\forall y$, from $P(\mathbf{x}|y,\theta) = P(x_1|y,\theta_1)\cdots P(x_d|y,\theta_d)$

$= \prod_{i=1}^{d} P(x_i, y, \theta_i)$, $\theta = (\theta_0, \ldots, \theta_d)^T$

and $\theta_0 = (P(y=1), \ldots, P(y=m))^T$

## Why Naive Bayes?

- Low Complexity

$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ → Binary feature (0/1), → Binary feature (0/1)

$O(md)$, relatively immune to overfitting

$P(\underbrace{x_1 \ldots x_d}_{x}|\,0)$ $\therefore 2^d$ distinct feature vectors

e.g. $d=2$ $\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}$ (4)

$d=3$ $\begin{pmatrix}0\\0\\0\end{pmatrix} \cdots \begin{pmatrix}1\\1\\1\end{pmatrix}$ (8)

$2^d - 1$ parameters to specify the joint pmf

# of params needed is growing exponentially fast

$O(md)$ vs. joint of $O(m(2^d-1))$

$\forall i, y$, $\hat{\theta}_{iy,ML}(D) = \arg\max_{\theta_{iy}} \left[\frac{1}{n_y} \sum_{j:\, y_j = y} \ln P(x_{ji}|y,\theta_{iy})\right]$

$= \hat{\theta}_{iy,ML}(D_{iy})$

## Gaussian Naive Bayes

$\forall i, y$, $P(x_i|y,\theta_i) \sim N(\mu_{iy}, \sigma^2_{iy})$

Categorical NB: All features are discrete

N.B. assumption: $P(\mathbf{x}_j|y_j) = \prod_{i=1}^{d} P(x_{ji}|y_j,\theta_i)$

## Categorical NB

Bayes Classifier for 0-1 loss (MPE rule = MAP rule)

$h_{MPE}(x) = \arg\max_{y=1,\ldots,m} P(y) \prod_{i=1}^{d} \beta_{x_i,y,i}$

$\forall i, y, \hat{\theta}_{iy,ML}(D) = \arg\max_{\theta_{iy}} \left[\sum_{w=1}^{W} \left(\frac{n_{w,y,i}}{n_y}\right) \ln \beta_{w,y,i}\right]$

$n_y$: # class $y$ train exs

$n_{w,y,i}$: # class $y$ train exs where $i$th feature $= w$

$\forall w, y, i$, $\hat{\beta}_{w,y,i} = \frac{n_{w,y,i}}{n_y}$

notation

$y$ = class label $1,\ldots,m$

$j$ = sample index $1,\ldots,n$

$i$ = feature index $1,\ldots,d$

$w$ = feature value $1,\ldots,W$

$y$ = class label $\in \{1,\ldots,m\}$

$j$ = sample index $\in \{1,\ldots,n\}$

$i$ = feature (component index) $\in \{1,\ldots,d\}$

$\theta_i$ = parameters for class conditional pmfs/pdfs of feature $i$ across all classes

$\Theta_i = \{\theta_{iy}, y=1,\ldots,m\}$, where

$\theta_{iy}$ = parameters for class-conditional feature pmf/pdf of feature $i$ in class $y$ (i for class $y$) and $P(x_i|y,\theta_i) = p(x_i|y,\theta_{iy})$ for all $i, y$

$D = \{(x_1,y_1), \ldots, (x_n,y_n)\}$ (training set)

## Document Classification (2 classes; $m=2$)

Vocabulary size $= W = 3$ e.g. $\{a, b, c\}$

Training set

class 1 examples

| $x_1$ | $x_2$ |
|---|---|
| a | b |
| b | c |
| b | a |
| c | b |
| a | c |
| c | c |

$d=1$, length $= 7$ $\quad$ $d=2$, length $= 5$

class 2 examples

| | | |
|---|---|---|
| c | c | c |
| c | c | b |
| a | c | c |
| b | | a |
| c | | |

$L=5$ $\quad$ $L=3$ $\quad$ $L=4$

likelihoods $P(\mathbf{x}|y)$

$\hat{\beta}_{a,1} = \frac{2+1}{7+5} = \frac{3}{12}$ ← # $a$'s, ← total words

$\hat{\beta}_{b,1} = \frac{2+2}{12} = \frac{4}{12}$

$\hat{\beta}_{c,1} = \frac{3+2}{12} = \frac{5}{12}$

$\hat{\beta}_{a,2} = \frac{2}{12}$

$\hat{\beta}_{b,2} = \frac{2}{12}$

$\hat{\beta}_{c,2} = \frac{8}{12}$

Priors

$\hat{p}(y=1) = \frac{2}{2+3} = \frac{2}{5}$, $\hat{p}(y=2) = \frac{3}{2+3} = \frac{3}{5}$

## Test document

NB assumption:

| $x$ | $P(\cdot)\,\vert_1$ | $P(\cdot)\,\vert_2$ |
|---|---|---|
| b | $P(b)\,\vert_1$ | $P(b)\,\vert_2$ |
| b | $P(b)\,\vert_1$ | $P(b)\,\vert_2$ |
| a | $P(a)\,\vert_1$ | $P(a)\,\vert_2$ |
| c | | |
| c | | |
| a | | |
| b | | |
| a | | |
| c | | |
| ∅ | | |
| c | | |
| c | | |

$d=12$

∅: don't mult. by 0 if $d$ not in train set OR can do a Bayesian Regularization (Laplace smoothing - default counts)

$P(\text{class } 1)$ $\quad$ $P(\text{class } 2)$

$\hat{p}(y=1) \cdot \hat{p}(\underset{\text{naive bayes}}{\underline{x}}\,|\,y=1)$

$= \frac{2}{5} \cdot \hat{\beta}_{a,1}^3 \cdot \hat{\beta}_{b,1}^4 \cdot \hat{\beta}_{c,1}^3$

then compare with

$\hat{p}(y=2) \cdot \ldots$

$= \frac{3}{5} \cdot (\hat{\beta}_{a,2})^3 \cdot (\hat{\beta}_{b,2})^4 (\hat{\beta}_{c,2})^5$

## Dirichlet Prior

## Example: Binary classification, $y$ = (lifestyle) Peaceful / Stressed Binary (0,1)

Feature Vector

Values

$\underline{x} = \begin{pmatrix} \text{Income} \\ \text{charitable giving} \\ \text{Avg \# hrs sleep per day} \end{pmatrix}$ → Low/middle/high ... Ternary, say Categorical over $(0,1,2)$

→ No/Yes ... Binary, say Bernoulli over $(0,1)$

→ 2-10 ... Real value, say Gaussian (scalar)

$d=3$

Training data

| | $y=0$ (peaceful) | | | | | | | $y=1$ (stressed) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (7) | | | | | | | (7) | | | | | | |
| 1) | 2 | 2 | 2 | 1 | 0 | 0 | 1 | 0 | 2 | 2 | 2 | 1 | 1 | 0 |
| 2) | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3) | 9 | 8 | 6.5 | 5 | 4 | 3.5 | 7 | 4 | 5 | 7 | 4 | 8 | 6.5 | 3 |

## Naive Bayes

$\theta_{0y} = p(y)$ $\quad$ $\hat{p}_{ML}(0) = \frac{7}{14} \leftarrow \frac{n_0}{n} = \hat{p}_{ML}(1)$ (sample mean)

First feature:

$\theta_{1y}$ = pmf over $(0,1,2)$

for class '0' $q_0, q_1, q_2$

for class '1' $r_0, r_1, r_2$

$y=0$: $\hat{q}_{0\,ML} \overset{?}{=} \frac{2}{7}$ | $\hat{q}_{1ML} \overset{?}{=} \frac{3}{7}$ | $\hat{q}_{2ML} \overset{?}{=} \frac{3}{7}$

$y=1$: $\hat{r}_{0ML} \overset{?}{=} \frac{2}{7}$ | $\hat{r}_{1ML} \overset{?}{=} \frac{2}{7}$ | $\hat{r}_{2ML} \overset{?}{=} \frac{3}{7}$

Second feature:

$\theta_{2y}$ (charity) = pmf over $(0,1)$

$\hat{\theta}_{0\,ML} = \left(\frac{2}{7}, \frac{5}{7}\right)$ (class 0 peaceful) $\quad$ $\hat{\theta}_{1\,ML} = \left(\frac{6}{7}, \frac{1}{7}\right)$ (class 1 stressed)

Third feature:

$\theta_{3y}$ (sleep) = for $y=0$, $\theta_{30} = (\mu_0, \sigma_0^2)$

for $y=1$, $\theta_{31} = (\mu_1, \sigma_1^2)$

estimated params.

$\hat{\mu}_{0ML} = \frac{9+8+6.5+\ldots}{7}$

$\hat{\sigma}^2_{0ML} = \ldots$

$\hat{\mu}_{1ML} = \ldots$

$\hat{\sigma}^2_{1ML} = \ldots$

Test individual

$\underline{x}_{test} = \begin{pmatrix}1\\1\\3\end{pmatrix}$

$y = ?$

Compare

MAP → ML b.c. $\frac{1}{2}, \frac{1}{2}$ in this case

$y=0$ (p(y)) $\quad \frac{1}{2} \cdot \frac{2}{7} \cdot \frac{5}{7} \cdot N(\hat{\mu}_{0ML}, \hat{\sigma}^2_{0ML})$ evaluated at 3

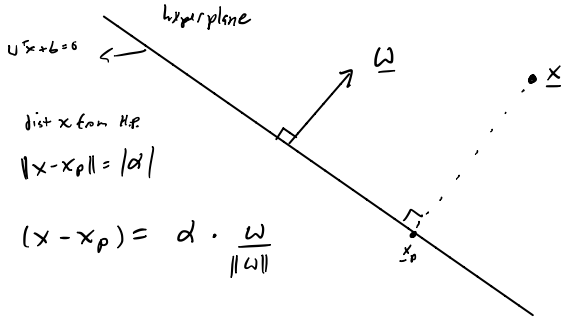
$y=1$ $\quad \frac{1}{2} \cdot \frac{3}{7} \cdot \frac{1}{7} \cdot N(\hat{\mu}_{1ML}, \hat{\sigma}^2_{1ML})$ (3)

$\gtrless$ ($\geq$ for $y=0$, $<$ for $y=1$)

# SVM: Find max margin hyperplane (convex optimization problem)

Equation of hyperplane in $\mathbb{R}^d$: $w^T x + b = 0$

$(w, b)$ and $(\gamma w, \gamma b)$ define the same hyperplane if $\gamma \neq 0$ — $(\gamma w)^T x + \gamma b = 0$, $\gamma \neq 0$




$(x - x_p) = \alpha \cdot \frac{w}{\|w\|}$

$\perp$ dist. (euclid.) of pt. $x$ from hyperplane $= \frac{|w^T x + b|}{\|w\|}$ ← euclidean

## Two-class Linearly separable SVM

- Label space $Y = \{-1, +1\}$
- Feature vectors $X = \mathbb{R}^d$
- Training set $D = \{(x_1, y_1), \ldots (x_n, y_n)\}$
- Family of classifiers: "linear" classifiers

$$H = \{h(x) = \text{sign}(w^T x + b) \quad w \in \mathbb{R}^d, \ b \in \mathbb{R}\}$$

- Parameters

$$\theta = \begin{pmatrix} w \\ b \end{pmatrix} \quad (d+1) \text{ scalar parameters}$$

$w^T x_p + b = 0$

$\Rightarrow w^T (x - x_p) = \alpha \frac{w^T w}{\|w\|} = \alpha \frac{\|w\|^2}{\|w\|} = \alpha \|w\|$

$\Rightarrow \alpha = \frac{w^T w - w^T x_p}{\|w\|} = \frac{w^T x + b}{\|w\|} \Rightarrow |\alpha| = \frac{|w^T x + b|}{\|w\|}$

- Geometric margin of a hyperplane $(w, b)$ wrt $D$

$$\rho(w, b) := \min_{1 \leq j \leq n} \frac{|w^T x_j + b|}{\|w\|}$$

(distance of feature vector which is closest to the hyperplane)

- Linearly separable training set $D$:

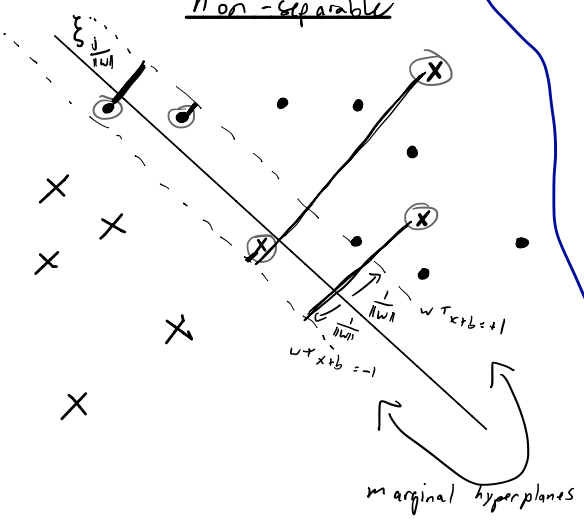
$$\exists_{\text{exists}} (w, b) \text{ s.t. } y_j (w^T x_j + b) > 0 \quad \forall j = 1, \ldots, n$$

$$\Rightarrow \boxed{\rho(w, b) > 0}$$

Canonical Parameterization of hyperplane $(w, b)$ wrt lin. sep. $D$:
If $D$ is lin. sep. by $(w, b)$ and $\rho > 0$, then the canonical parameterization $(w_{canon}, b_{canon})$ is such that

$$\min_{1 \leq j \leq n} |w_{canon}^T x_j + b_{canon}| = 1$$

## Non-separable




$y_j (w^T x_j + b) \geq 1 - \xi_j$ ← slack for $j$

objective function of SVM:

$$(w_{svm}, b_{svm}) = \underset{(w,b):}{\text{argmin}} \left[ \frac{1}{2} \|w\|^2 + \text{penalty}(\xi_1, \ldots, \xi_n) \right]$$

$$\forall j \quad y_j (w^T x_j + b) \geq 1 - \xi_j, \quad \xi_j \geq 0$$

$\{x_j : \alpha_j > 0\} \rightarrow$ Support Vectors

$\{x_j : \xi_j > 0\} \rightarrow$ Outliers

All outliers are support vectors
→ from complementary slackness condition #2:
$\xi_j^* > 0 \Rightarrow \alpha_j^* = C_j > 0$
Support vectors $\Rightarrow \alpha_j^* \neq 0$

$$(w_{svm}, b_{svm}) = \underset{(w,b):}{\text{argmax}} \frac{1}{\|w\|}$$

1) $y_j (w^T x_j + b) > 0 \quad \forall j$
2) $\min_{1 \leq j \leq n} |w^T x_j + b| = 1$

$$\rho_{opt} = \rho(w_{svm}, b_{svm}) = \max_{(w,b)} \frac{1}{\|w\|}$$